# MULTIJEPA: LEARNING MUSIC REPRESENTATIONS ACROSS TEMPORAL RESOLUTIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We present **MultiJEPA**, a multiresolution joint-embedding predictive architecture for learning representations of signals whose semantics occur at multiple resolutions. Rather than committing to a single resolution, MultiJEPA trains one encoder across dynamically aggregated resolutions. In music, semantics span timescales from the millisecond-scale precision required for onset and beat localization to harmonic context and musical form unfolding over tens of seconds to minutes. Music is therefore an attractive domain for studying how temporal resolution relates to learned semantics. Specifically, a controlled five-resolution analysis reveals a strong context–detail tradeoff: coarse representations favor tonal and structural understanding, while fine representations favor beat tracking and transcription, with no single resolution dominating. MultiJEPA broadly recovers these fixed-resolution profiles with a single encoder queried at four temporal rates. Simple four-query fusion exploits complementary information across resolutions, outperforming all compared audio-only SSL baselines on 15 of 16 probing metrics spanning nine downstream tasks. We further propose a short fine-to-coarse self-alignment phase that organizes these views into a nested representation and improves cross-resolution probe transfer. Together, these results establish temporal resolution as an explicit design axis for music understanding models, and show that MultiJEPA pushes the Pareto frontier of the context-detail tradeoff.

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.

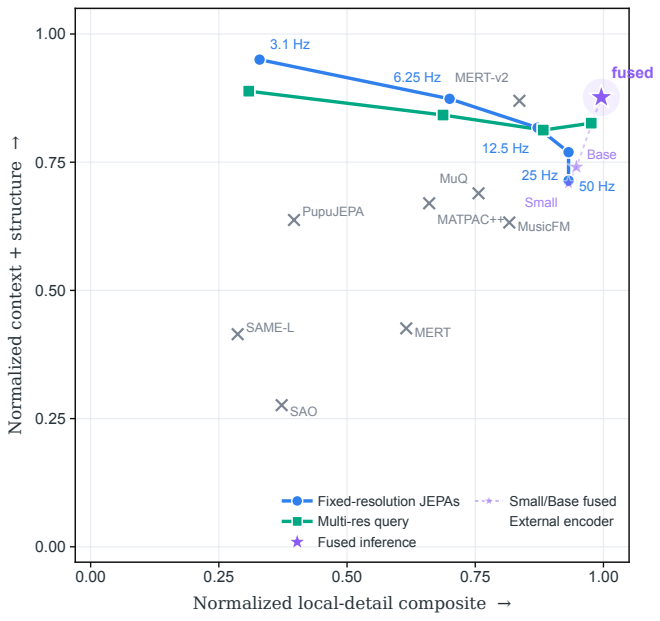


Figure 1: **MultiJEPA** extends the context–detail frontier under a composite of coarse and fine tasks. The horizontal composite captures local detail (beat, transcription, F0); the vertical composite captures context and structure (genre, key, structure, tagging). Fixed-resolution JEPAs and multiresolution queries trace similar frontiers, while four-query fusion moves beyond both. Gray crosses denote external encoders and autoencoder references; small stars show fused Small and Base models.

# 1 INTRODUCTION

Self-supervised representation learning has become the de facto approach to training generalizable semantic backbones, seeing wide success across image (Oquab et al., 2024; Siméoni et al., 2025; Assran et al., 2023), video (Bardes et al., 2024; Assran et al., 2025; Mur-Labadia et al., 2026), and increasingly music (Li et al., 2024; Zhu et al., 2025; Won et al., 2024; Fei et al., 2023; Tuncay et al., 2025; Quelennec et al., 2025b). These systems, which in themselves seek only to learn "useful representations," form the backbone for an increasingly diverse range of downstream tasks, from genre classification and beat tracking (Li et al., 2024; Won et al., 2024), to similarity-based retrieval (Akama et al., 2023; Zhu et al., 2025), to conditioning for generation (Rouard et al., 2024; Xu et al., 2026; Lin et al., 2026). Among music models, joint-embedding predictive architectures (JEPAs Quelennec et al. (2025b); Fei et al. (2023); Gu et al. (2026)) and tokenized masked prediction have become common modeling choices, with models like MuQ, MATPAC and MusicFM showing strong performance in a range of downstream probing tasks (Zhu et al., 2025; Won et al., 2024).

Despite this progress, many widely used music encoders still make two simplifying choices. First, most are trained on excerpts ranging from 5–30 seconds (Li et al., 2024; Zhu et al., 2025; Won et al., 2024), despite the fact that musical context extends to minute-scale tracks. This potentially limits their applicability to long-context tasks such as structural analysis, retrieval, or generation, where broader context can matter. Second, these approaches learn representations at a *single temporal resolution*. This becomes particularly restrictive for music, as semantic events range from tens of milliseconds (note onsets and beats) to minutes (musical form and harmonic context), and the semantics at these scales can be quite different. Additionally, at full-track length, the gap between these orders of magnitude makes a single fine-grained representation computationally difficult, introducing an engineering tradeoff between context length and representation resolution. Given the wide range of scales in which music semantics arise, a general-purpose representation should therefore be able to represent both coarse, long-term structure and short, high-resolution events.

In this work, we propose **MultiJEPA**, a new representation learning framework extending JEPA to multiple resolutions. To do so, MultiJEPA learns representations at several resolutions within a shared encoder, internalizing semantics native to each one. We then further propose a short *cross-resolution self-alignment* phase, which aligns the representations of different resolutions in a nested manner and makes fine and pooled coarse representations available in a single inference call. Across our 9-task probing suite, our fixed- and multiresolution models achieve the best result among the compared audio-only SSL baselines in 15 of 16 metrics. Notably, equal-weight fusion of four native queries, interpolated to the finest grid, outperforms the finest resolution on several fine-grained tasks, showing that different scales contain complementary information. Our key contributions are:

1. MultiJEPA, a SOTA music SSL system with independently queryable resolutions in one encoder and a short self-alignment phase matching pooled fine features to native coarse features.
2. A parameter-free fusion mechanism which extends the observed Pareto frontier and surpasses the compared audio-only SSL baselines in 15 of 16 coarse- and fine-grained probing metrics.
3. A controlled five-resolution analysis, which exposes a semantic tradeoff: fine representations preserve beat and transcription detail but cover less musical context, while coarse representations favor key and structure. No fixed temporal rate dominates the task suite.

# 2 MULTIJEPA: MULTIRESOLUTION JOINT-EMBEDDING PREDICTIVE ARCHITECTURE

MultiJEPA extends masked latent prediction on mel-spectrograms (Sec. 2.1) with a shared multiresolution patchifier (Sec. 2.2). A short self-alignment phase then matches pooled fine features to their native coarse counterparts (Sec. 2.3) ( See Appendix A for more details).

## 2.1 DEEP JEPA ON MEL-SPECTROGRAMS

We train MultiJEPA with a masked latent prediction objective in a JEPA architecture (Assran et al., 2023; Bardes et al., 2024; Mur-Labadia et al., 2026). A linear patchifier of size $(p_T, p_F)$ maps an input mel-spectrogram $X_a$ to a time-frequency token sequence $z = \text{Patchify}(X_a)$. We sample

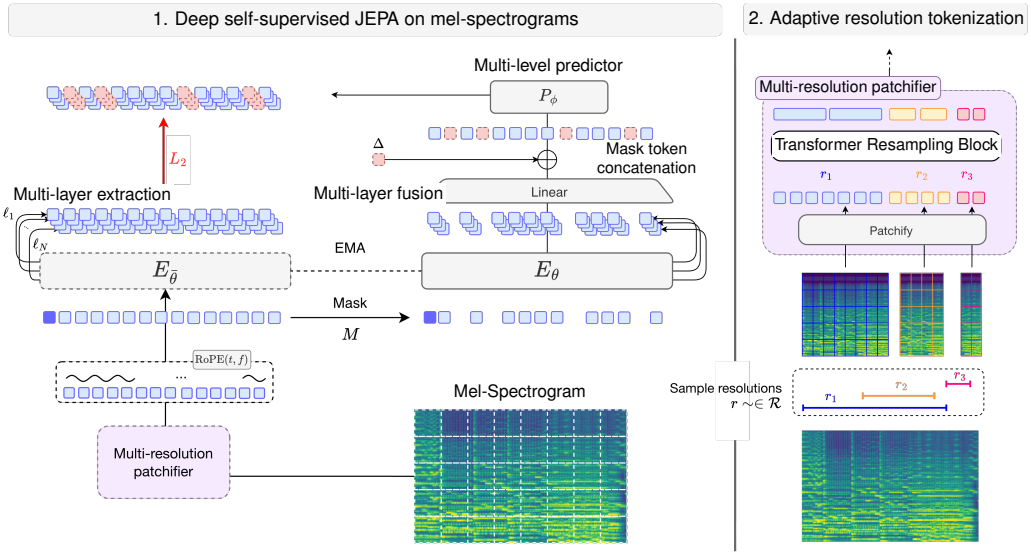


Figure 2: **MultiJEPA architecture. Left:** a masked student predicts stop-gradient features from an EMA teacher at four encoder depths. **Right:** each training step samples a temporal resolution and its associated crop duration. A common fine patch grid is mapped by a Transformer Resampling Block to the same $K$-token budget at every rate, so coarse queries cover longer spans while fine queries preserve denser timing. The patchifier, encoder, predictor, and teacher are shared across rates.

masked indices $M$ and visible indices $C$ using the hybrid masking strategy in Table 5. The student processes only visible tokens $z_C = (z_i)_{i \in C}$, while the EMA teacher processes the full sequence.

A predictor uses the visible student features and learned mask tokens to predict the teacher features. Let $\hat{y}$ denote its predictions and $\bar{y}$ the teacher targets.

We minimize the squared error at masked positions:

$$\mathcal{L} = \frac{1}{|M|} \sum_{i \in M} \|\hat{y}_i - \mathrm{sg}(\bar{y}_i)\|_2^2, \tag{1}$$

where sg denotes stop-gradient. The teacher targets distill contextual semantics into the student through information asymmetry. We update the teacher by EMA, $\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta$.

V-JEPA 2.1 (Mur-Labadia et al., 2026) shows that multi-level supervision improves representation quality. We extend this approach to music with deep SSL, predicting teacher representations at four encoder depths. Appendix A details deep SSL, and Appendix D reports the comparison with final-layer-only supervision.

### 2.2 MULTIRESOLUTION TRAINING

Single-resolution JEPAs (Sec. 2.1) learn one operating point on the context–detail tradeoff. In contrast, **MultiJEPA** learns several operating points by training across multiple resolutions. We define a set of *resolutions* $\mathcal{R}$, where each resolution $r \in \mathcal{R}$ has an effective patch size $(p_T^r, p_F^r)$ and crop length $L^r$ (in spectrogram frames). We fix a constant $p_F$ for all resolutions*. All resolutions are patchified at the finest patch size $p_T^f$. To construct inputs at a given resolution, we pool the resulting fine-resolution tokens to the corresponding frame rate. At each frequency position, the pooling module $\mathrm{Pool}_r : \mathbb{R}^{a_r \times d} \to \mathbb{R}^d$ aggregates $a_r = p_T^r / p_T^f$ consecutive fine tokens into one coarse token before masking. The encoder then processes the pooled tokens using the standard JEPA objective described in Section 2.1.

*Varying frequency resolution could expose a different tradeoff for tasks such as transcription; we leave this for future work.

We pool *before* the encoder to avoid computational burden on longer crops. This allows us to train at coarser resolutions with longer contexts. To avoid unpredictable attention cost fluctuations during training, we fix a constant token length for all resolutions: $K = \frac{L^r}{p_T^r} \times \frac{F}{p_F^r}$, with $K = 2048$, and $F$ the number of mel bins. Thus, each operating point pairs a temporal resolution with a corresponding crop duration while keeping the encoder token budget fixed (Table 1). The encoder therefore receives the same sequence length and masking distribution at every operating point, while each output token summarizes a rate-dependent span of audio. In the semantic encoder, continuous 2D RoPE (Su et al., 2024) assigns positions from each token's physical center time and frequency: The same location receives the same positional signal across rates.

Table 1: Temporal resolution operating points.

| **Patch** | **Crop** (frames) | **Dur.** (s) | Rate (**Hz**) |
|---|---|---|---|
| $32 \times 16$ | 8192 | ∼82 s | 3.1 |
| $16 \times 16$ | 4096 | ∼41 s | 6.25 |
| $8 \times 16$ | 2048 | ∼20 s | 12.5 |
| $4 \times 16$ | 1024 | ∼10 s | 25.0 |
| $2 \times 16$ | 512 | ∼5 s | 50.0 |

Our pooling operator is a lightweight **T**ransformer **R**esampling **B**lock (Parker et al., 2026). For each local group of $a_r$ fine tokens at a fixed frequency position, the TRB appends a learned query and applies block-local self-attention. Only the query output is retained, reducing the sequence length by a factor of $a_r$; the finest resolution bypasses the TRB. Unlike the original implementation, our TRB uses no RoPE; its block sliding-window attention provides a locality bias. TRB supports additional resolutions without requiring new resolution-specific patchifiers, as used in CrossViT (Chen et al., 2021). Its attention-based aggregation also adapts naturally across resolutions, avoiding the need to resize learned patch-embedding weights as in FlexiViT (Beyer et al., 2023). Unlike convolutional pooling, the TRB's pure attention aggregation is simpler to adapt to multiple resolutions without multiple patchifiers. Appendix A provides more details on the TRB architecture.

Finally, motivated by the expectation that different resolutions capture different and complementary semantics, we evaluate a combination of all resolutions (*fusion*): We extract representations at each $r$, linearly interpolate each one to the finest time grid, and average them.

### 2.3 Cross-Resolution Alignment

Training at multiple resolutions learns scale-specific and potentially complementary semantics, but does not explicitly encourage consistency across views. We hypothesize that cross-resolution alignment can encourage more consistent semantic structure across scales and enable coarse features to be recovered from a single fine-resolution forward pass. We therefore add a short **cross-resolution alignment** phase after pretraining. Alongside the ongoing MultiJEPA objective, we sample coarse ($r_t$) and fine ($r_s$) resolutions. Both views process the same audio crop using the crop duration associated with $r_t$. The teacher processes the crop at $r_t$, while the student processes it at $r_s$. To encourage fine representations to contain local variation around a shared coarse component, we block-average the student features onto the coarse grid, yielding $K_c$ tokens aligned with the corresponding teacher features.

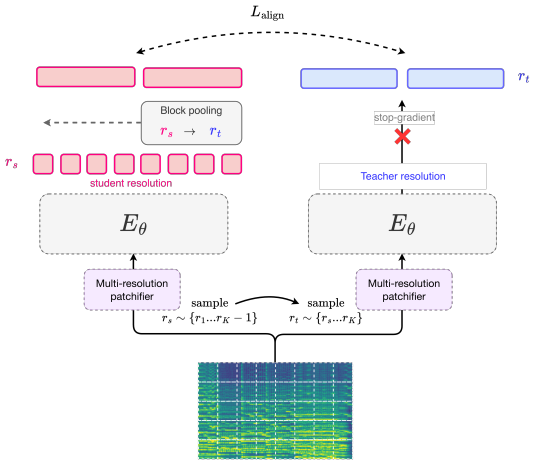


Figure 3: Self-alignment block-averages a fine query onto the native coarse grid and matches the stop-gradient coarse view. Only the fine student branch receives the alignment gradient.

Let $q$ denote the integer temporal-grid ratio between $r_s$ and $r_t$. The fixed map $\rho_{f \to c}$ averages each disjoint group of $q$ fine tokens corresponding to one coarse token. This yields an exact decomposition into a block mean and a zero-mean within-block residual. Let $Y_s$ denote the student features at $r_s$, $\rho(\cdot)$ the block-pooling operator, $d$ the feature dimension, and $Y_c$ the EMA teacher features at $r_c$.

$$\mathcal{L}_{\text{feat}} = \frac{1}{K_c d} \sum_{i=1}^{K_c} \left\| \rho(Y_s)_i - \text{sg}\big((Y_c)_i\big) \right\|_2^2 \tag{2}$$

The total alignment-phase loss is $\mathcal{L} = \mathcal{L}_{\text{deep}} + \lambda_{\text{feat}} \mathcal{L}_{\text{feat}}$, where $\mathcal{L}_{\text{deep}}$ is the prediction loss over the supervised depths. We set $\lambda_{\text{feat}} = 1$, which we found to work well empirically, and leave a systematic study of the alignment weight to future work.

After alignment, a single fine-resolution forward pass yields both fine features and provides a single-pass approximation to native coarse representations.

# 3 Experiments

## 3.1 Experimental Setup

All models are trained on approximately 250k hours of an internal dataset. Training samples a random 5–82 s crop, depending on the operating point; the encoder is a 24-layer, 1024-dimensional Vision Transformer with continuous 2D RoPE (Dosovitskiy et al., 2021; Su et al., 2024). We train it for 400k steps with AdamW, hybrid masking, deep supervision at four depths $\{L/4, L/2, 3L/4, L\}$, and an effective batch size of 1024. The multiresolution large model has approximately 420M parameters, including the TRB. Appendix A gives the complete optimization and architecture specifications. In addition to MultiJEPA, we train simple single-resolution Deep JEPAs (Section 2.1) at the operating points defined in Table 1. This allows us to (1) show the coarse-fine tradeoff at single-resolution training points and (2) assess whether the shared model reproduces the fixed-resolution models' rate-dependent performance patterns when switching to multiresolution training.

We compare representations on frozen linear downstream probing (See Section 3.2) against MERT (Li et al., 2024), MERT-v2, MuQ (Zhu et al., 2025), MusicFM (Won et al., 2024), MATPAC++ (Quelennec et al., 2025b), and PupuJEPA (Gu et al., 2026). To broaden the field, we include results from probing two autoencoders – the semantically-aligned SAME-L (Parker et al., 2026) and the Stable Audio Open (SAO) autoencoder (Evans et al., 2024) – as well as MERT-v2. MERT-v2 predicts masked codes derived from MuQ and Qwen2-Audio features; the latter is a 7B model trained on paired audio and text (Chu et al., 2024). We include MERT-v2 as a state-of-the-art music representation baseline, but distinguish its inherited weakly-supervised audio–language distillation from audio-only SSL in Table 2.

## 3.2 Evaluation Datasets and Protocol

We evaluate global semantics, temporal detail, and structure using the tasks in Table 6 (Appendix B). As is standard in SSL literature, our probes are deliberately simple to measure representation quality rather than the task performance ceiling. Clip-level probes train and evaluate on window-averaged embeddings for each dataset recording. For the main probing suite, we select one encoder layer per checkpoint using intrinsic dimension (ID) (Kanatas et al., 2026) and use it for every downstream task; The same is done for baselines. Downstream probes are early-stopped on validation performance. Reported scores are averaged over five randomly seeded probe runs. Appendix B records the remaining protocol details.

The encoder is frozen and the EMA teacher supplies representations unless noted. Frame tasks stitch window features and resample them to the label grid. multiresolution queries are evaluated separately. For structure, a relational projection precedes clustering; zero-shot segmentation uses the same clustering directly on frozen frames (Appendix B.2).

## 3.3 Strong Music Representations Across Temporal Scales

MultiJEPA surpasses the strongest compared audio-only SSL baseline on 15 of 16 metrics with four-query fusion (Table 2). The single-resolution results reveal why combining scales is useful: coarse representations favor key and structure, while finer representations improve beat tracking and transcription. Genre and tagging vary much less with resolution. Between 3.1 and 50 Hz, Harmonix relational mIoU falls from 49.0 to 39.3, while Beat F1 at 70 ms rises from 40.0 to 90.5. These opposing trends expose the context–detail tradeoff: no fixed temporal rate dominates the task suite.

We interpret these scores in relation to resolution. At 3.1 Hz, temporal positions are roughly 320 ms apart, making high-resolution events such as 20ms tolerance beats unable to be sharply located. Beat and onset localization deteriorate sharply under our probing protocol with lower resolutions.

Table 2: Linear probing across semantic, tonal, structural, rhythmic, transcription, and F0 tasks; scores are five-probe-run means. Fixed-resolution rows use separate encoders; all four native-query rows share one multiresolution encoder. Bold values mark the best results in each column. Underlined values mark metrics where parameter-free four-query fusion surpasses every audio-only SSL baseline. MERT-v2 inherits audio–language supervision through its Qwen2-Audio targets.

| | | | GTZAN | MTAT | | OpenMIC | FMAK | GiantSteps | Structure | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Rate (Hz) | Context (s) | Acc. | AUROC | mAP | mAP | MIREX | MIREX | Rel. mIoU | Rel. Pair F | ZS mIoU | ZS Pair F |
| *External encoders* | | | | | | | | | | | | |
| MERT | 75 | 5 | 66.2 | 84.0 | 31.3 | 51.4 | 61.0 | 54.0 | 34.6 | 52.3 | 22.3 | 42.7 |
| MuQ | 25 | 10 | 82.1 | 91.6 | 46.6 | 67.5 | 60.0 | 57.2 | 40.8 | 61.8 | 17.9 | 41.5 |
| MusicFM | 25 | 30 | 76.8 | 90.9 | 44.6 | 61.9 | 60.9 | 58.8 | 38.3 | 57.5 | 18.4 | 39.4 |
| MATPAC++ | 6.25 | 6 | 86.6 | 91.7 | 47.0 | 75.5 | 63.4 | 57.7 | 37.5 | 58.0 | 30.7 | 50.8 |
| PupuJEPA | 25 | 10 | 84.6 | 92.1 | 48.4 | 75.5 | 63.2 | 55.2 | 41.8 | 61.5 | 27.4 | 48.5 |
| *VAEs and supervised distillation* | | | | | | | | | | | | |
| MERT-v2 | 25 | 360 | **91.0** | 92.2 | 48.5 | 73.2 | 67.3 | 62.7 | 43.2 | 61.3 | 19.2 | 32.4 |
| SAO | 21.5 | — | 41.4 | 80.2 | 26.8 | 31.1 | 63.4 | 61.6 | 29.6 | 45.6 | 21.3 | 40.9 |
| SAME-L | 12 | — | 50.3 | 81.6 | 29.1 | 37.0 | 66.5 | **65.6** | 30.7 | 49.6 | 22.7 | 45.3 |
| *MultiJEPA: fixed-resolution JEPAs* | | | | | | | | | | | | |
| 32 × 16 | 3.1 | 82 | 84.8 | 91.8 | 47.9 | 66.1 | 68.0 | 64.3 | **49.0** | **68.0** | **39.7** | **58.1** |
| 16 × 16 | 6.25 | 41 | 88.6 | 92.3 | **49.1** | 71.5 | **69.3** | 62.2 | 44.8 | 64.8 | 35.8 | 53.9 |
| 8 × 16 | 12.5 | 20 | 89.0 | **92.3** | 48.9 | 72.5 | 66.6 | 61.2 | 42.2 | 61.6 | 31.2 | 49.6 |
| 4 × 16 | 25 | 10 | 89.3 | 92.2 | 48.6 | 74.6 | 65.5 | 58.5 | 41.5 | 60.1 | 21.9 | 46.9 |
| 2 × 16 | 50 | 5 | 87.6 | 92.2 | 48.7 | **75.7** | 64.9 | 58.1 | 39.3 | 58.0 | 20.8 | 47.7 |
| *MultiJEPA: multiresolution JEPA* | | | | | | | | | | | | |
| 32 × 16 | 3.1 | 82 | 87.2 | 91.7 | 47.6 | 64.2 | 61.8 | 61.5 | 46.8 | 65.6 | 39.4 | 57.0 |
| 16 × 16 | 6.25 | 41 | 88.3 | 92.1 | 48.0 | 69.3 | 65.4 | 60.6 | 44.4 | 64.3 | 36.5 | 55.8 |
| 8 × 16 | 12.5 | 20 | 87.2 | 92.0 | 48.2 | 70.6 | 62.8 | 60.4 | 43.3 | 62.8 | 28.4 | 50.0 |
| 4 × 16 | 25 | 10 | 87.2 | 91.9 | 48.2 | 72.6 | 65.1 | 61.7 | 43.0 | 61.2 | 25.3 | 47.2 |
| **Four-query fusion** | — | 10–82 | 87.6 | 92.2 | 48.7 | 71.8 | 64.9 | 60.9 | 46.4 | 66.3 | 36.2 | 55.4 |

*Fine-grained tasks; transcription averages GuitarSet and MusicNet-EM, and F0 reports Melody2 RPA and Multi-F0 F1.*

| | | | Rhythm | | Transcription | | F0 | |
|---|---|---|---|---|---|---|---|---|
| Model | Rate (Hz) | Context (s) | $F1_{70}$ | $F1_{20}$ | Frame F1 | Onset F1 | RPA | Multi-F0 F1 |
| *External encoders* | | | | | | | | |
| MERT | 75 | 5 | 87.3 | 65.5 | 57.2 | 33.3 | 26.6 | 41.5 |
| MuQ | 25 | 10 | 88.6 | 56.3 | 66.4 | 47.4 | 46.1 | 41.0 |
| MusicFM | 25 | 30 | 86.8 | 49.4 | 67.3 | 50.7 | 54.6 | 47.2 |
| MATPAC++ | 6.25 | 6 | 78.9 | 40.0 | 69.6 | 34.2 | 38.6 | 47.4 |
| PupuJEPA | 25 | 10 | 84.2 | 49.5 | 56.8 | 34.2 | 16.1 | 22.8 |
| *VAEs and supervised distillation* | | | | | | | | |
| MERT-v2 | 25 | 360 | 89.5 | 67.1 | 71.6 | 45.0 | 40.7 | 45.2 |
| SAO | 21.5 | — | 70.6 | 45.6 | 44.2 | 27.5 | 31.0 | 33.2 |
| SAME-L | 12 | — | 52.1 | 22.9 | 49.3 | 26.0 | 39.2 | 32.1 |
| *MultiJEPA: fixed-resolution JEPAs* | | | | | | | | |
| 32 × 16 | 3.1 | 82 | 40.0 | 15.6 | 65.4 | 11.3 | 50.5 | 36.2 |
| 16 × 16 | 6.25 | 41 | 81.5 | 45.2 | 69.4 | 27.6 | 56.6 | 43.0 |
| 8 × 16 | 12.5 | 20 | 88.9 | 61.8 | 73.6 | 50.7 | 54.4 | 43.6 |
| 4 × 16 | 25 | 10 | 89.9 | 67.3 | 75.2 | 58.6 | 55.0 | 44.0 |
| 2 × 16 | 50 | 5 | **90.5** | 67.0 | **75.4** | **61.0** | 52.8 | 43.7 |
| *MultiJEPA: multiresolution JEPA* | | | | | | | | |
| 32 × 16 | 3.1 | 82 | 39.5 | 15.4 | 64.6 | 12.3 | 45.3 | 36.4 |
| 16 × 16 | 6.25 | 41 | 81.9 | 45.0 | 69.9 | 27.8 | 53.1 | 42.4 |
| 8 × 16 | 12.5 | 20 | 89.1 | 61.9 | 73.0 | 50.3 | 55.9 | 45.0 |
| 4 × 16 | 25 | 10 | 90.4 | **68.8** | 75.3 | 59.5 | 58.7 | 47.2 |
| **Four-query fusion** | — | 10–82 | 90.1 | 68.7 | 75.3 | 60.4 | **61.1** | **48.6** |

Conversely, at 50 Hz, the same 2048-token training budget covers only about five seconds. For tasks requiring long-range musical context, such as key and structure, the representations cannot contextualize longer-horizon events. Frame transcription changes more gradually because pitch activity can survive after onset timing is blurred.

**One encoder broadly recovers the rate-dependent profiles.** Although each query receives only a fraction of the training updates and shares all backbone parameters, the multiresolution encoder broadly reproduces the fixed-resolution profiles, with gaps at the coarse end on key and structure. Its 3.1-Hz query reaches 46.8 Harmonix mIoU, retaining most of the coarse advantage, while its finer queries recover the beat and transcription endpoints. Figure 4 shows the raw curves, and Figure 1 shows that the queries cover much of the fixed-resolution frontier under the stated task

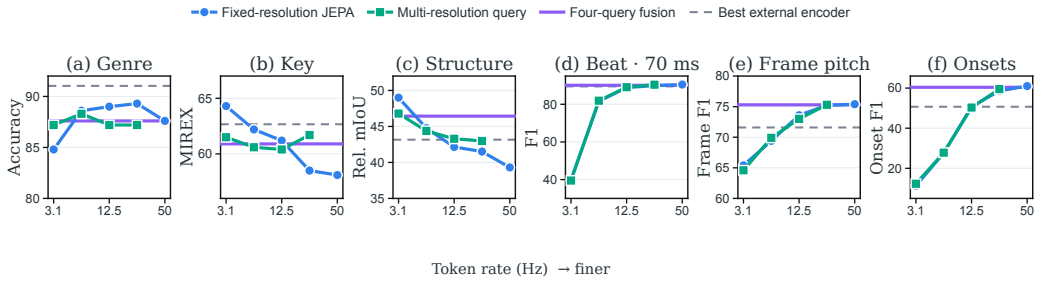


Figure 4: **Fixed-budget operating points and multiresolution fusion**. **(a–c)** Genre varies little with rate, while key and structural segmentation favor longer-context coarse views. **(d–f)** Beat tracking and transcription improve toward finer rates. Parameter-free fusion approaches the best resolution on the local tasks while retaining most of the coarse performance and structural score.

composite rather than collapsing toward one average representation. Interestingly, the clear trend for key estimation seems to break under multiresolution training, with 25Hz representations winning over coarser resolutions.

**Fusion exploits complementary scales.** Equal-weight fusion is within 0.4 points of the best query on Beat $F1_{70}$ and relational mIoU, improves both F0 metrics and reaches the strongest local-detail composite. Together with the 3.1-Hz fixed-resolution JEPA, it forms the nondominated set under the proposed pareto tradeoff. The composite averages min–max-normalized beat, transcription, and F0 for local detail, and genre, tagging, key, and relational mIoU for context and structure; Table 2 and Figure 4 show every metric. Genre and key show no clear winner, so the individual queries remain useful even where fusion is strongest overall. Across the fixed-resolution models and four-query fusion, MultiJEPA leads the compared external SSL encoders on tagging, key estimation, beat detection, transcription, pitch estimation, and structure. VAEs are notably strong on key estimation: SAME-L is strongest on GiantSteps, as its training includes explicit chroma-regression targets. Previous work has established that the best performing representations for key estimation fluctuate with layers and tasks (Li et al., 2024; Castellon et al., 2021). We encourage future work to understand and characterise this fluctuation in light of these new results with VAEs. SAME-L and SAO are weaker than the best MultiJEPA rate on all other tasks.

**Inference context and context generalization** The coarsest fixed-resolution JEPA reaches 49.0 relational mIoU and 68.0 pairwise F, improving over the strongest external audio-only SSL baselines by 7.2 and 6.2 points. The same ordering appears without a learned projection: the coarsest fixed and multiresolution views reach 39.7 and 39.4 zero-shot mIoU, versus 30.7 for MATPAC++.

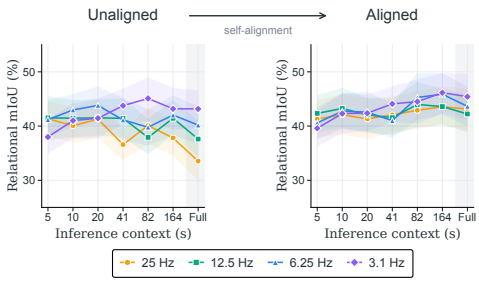


Figure 5: Structure probing before and after self-alignment. Probes are fitted at each model, rate, and context using the ID-selected encoder layer. Bands: 95% track-bootstrap intervals.

To evaluate RoPE-enabled context generalization and the effect of self-alignment, We vary inference context for the shared encoder before and after self-alignment in Figure 5, fitting each probe on features extracted at the same context used for evaluation. Every condition covers the full track. At a given inference context, the queries see identical audio with different token counts. Before alignment, the 3.1-Hz query rises from 38.0 relational mIoU at 5.12 s to 45.1 at 81.92 s and retains 43.2 with full-track inference. The 25-Hz query instead changes from 41.4 to 40.1 and then 33.5. Coarse queries retain more of their structural performance as inference extends beyond the training crop.

After self-alignment, all four queries reach 42.2–45.4 mIoU with full-track inference: the spread across rates narrows from 9.6 to 3.2 points. The 25-Hz query now retains its 81.92-s performance at full-track length (42.9 versus 43.1), while the 3.1-Hz query reaches 45.4. The phase improves

long-context structural segmentation most at the fine end and reduces its dependence on the query rate. Appendix G gives the continued-pretraining controls, zero-shot curves, and boundary results. Alongside these quantitative results, we provide qualitative visualizations of the model's full-track representations and the context–detail tradeoff in Appendix C.

### 3.4 Cross-Resolution Alignment and Consistency

Unaligned rates are not mutually substitutable: pooled fine features disagree with native coarse features despite sharing encoder weights. Self-alignment samples all six fine-to-coarse pairs among the four rates and optimizes Equation 2. We measure cross-resolution geometry with Centered Kernel Nearest-Neighbor Alignment (CKNNA) (Huh et al., 2024). Here CKNNA compares paired real views of the same audio: it tests whether pooled fine student and native coarse EMA-teacher representations induce the same neighborhoods across track means. Appendix F.2 defines the estimator. Training diagnostics, native-probe trajectories, and all pairwise results appear in Appendix F; Appendix F.3 examines the temporal spectra of these representations.

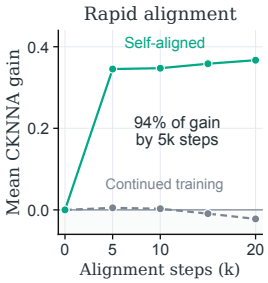


Figure 6: Mean CKNNA@10 gain from alignment start across all six rate pairs.

Self-alignment changes cross-resolution geometry almost immediately (Figure 6). Mean CKNNA@10 rises from 0.208 at the shared parent to 0.554 after 5k alignment steps and 0.575 after 20k; matched continued pretraining instead falls to 0.186. The first 5k steps account for 94% of the observed 20k gain. Every pair improves: after 20k steps, each aligned pair exceeds its matched control by 0.27–0.53 CKNNA, so the gain is not driven only by the easiest pairs.

For EMA results, alignment changes cross-resolution agreement much more than native geometry or probe scores (Appendix F).

Self-alignment also improves the fine query's long-context performance. With full-track inference and a probe fitted at that context, 25-Hz relational mIoU rises from 33.5 to 43.1, compared with 35.2 under continued pretraining. Boundary $F1_{0.5}$ rises from 27.9 to 47.9, versus 31.1 for the control. Zero-shot performance does not improve at every rate. Self-alignment improves cross-resolution organization and probe transfer, but leads to slight performance loss on fine tasks. Appendix G reports all context curves.




Figure 7: Probe transfer. Bars average retention over tasks per rate pair, before and after alignment. Curves show average retention for key, structure, and beat; dashed lines mark alignment start.

The measurements test complementary aspects of alignment. The blockwise loss measures temporal correspondence between pooled fine tokens and their coarse targets. CKNNA measures neighborhood agreement across track-level representations, while probe transfer tests whether a native-coarse head also works on pooled fine features. Geometric similarity alone does not necessarily preserve the directions used by a downstream probe.

Figure 7 reports pooled-fine performance as a percentage of native-coarse performance with the same validation-selected head; both views use the student. Most of the improvement arrives within 5k steps: mean key retention rises from 41.9% to 82.6%, structure from 85.3% to 94.9%, and beat from 71.1% to 87.6%. Key reaches 86.5% after 20k steps, compared with 42.1% under continued pretraining. For 25→12.5-Hz pooling, all seven tasks retain 99.2–100.9% of native-coarse performance after 30k steps. Pooling from 25 to 3.1 Hz still loses key and beat performance, so the coarse views are not fully interchangeable with pooled fine features. The bars include every task; Appendix F retains all seven trajectories.

## 4 Related Work

**Self-supervised music encoders.** Masked music encoders predict acoustic, codec, or random-projection targets (Li et al., 2024; Zhu et al., 2025; Won et al., 2024), while DINO and iBOT predict self-distilled teacher assignments in vision (Caron et al., 2021; Zhou et al., 2022; Oquab et al., 2024). JEPAs instead regress teacher features through a learned predictor (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025); supervising several depths improves the spatial quality of the learned features (Mur-Labadia et al., 2026). Music foundation models span both JEPAs/iBOT models (Quelennec et al., 2025a;b; Fei et al., 2023; Tuncay et al., 2025; Gu et al., 2026) and acoustic tokenization(Won et al., 2024; Zhu et al., 2025; Li et al., 2024). These models use one output rate and are generally trained on 5–30-second contexts. MultiJEPA applies JEPA-style self-distillation but makes temporal rate a training variable.

**Multiresolution training.** FlexiViT (Beyer et al., 2023) shares a vision transformer across patch sizes, trading sequence length for inference cost; FlexiAST (Feng et al., 2023) extends this approach to supervised audio classification. ElasticAST (Feng et al., 2024) trains across audio lengths and resolutions with supervised classification, rather than a self-supervised objective. In audio, Tailed U-Net (Vélez Vásquez & Burgoyne, 2022) combines musical timescales through a contrastively trained encoder–decoder, and MR-HuBERT (Shi et al., 2024) predicts masked speech units at several rates within a hierarchical encoder. To our knowledge, MultiJEPA is the first JEPA to jointly learn multiple queryable temporal resolutions of the same audio modality.

## 5 Discussion and Future Work

MultiJEPA shows that multiresolution training may benefit other music SSL paradigms, and more broadly, domains where semantics vary across scale (such as video). Several questions remain: the coarsest resolutions still underperform matched fixed-resolution models. We leave resolution extrapolation, dynamic resolutions, frequency resolution, and waveform-based JEPAs for future work.

Self-alignment leads to a small drop in fine-grained performance, notably transcription. V-JEPA 2.1 similarly trades classification accuracy for improved dense features before recovering performance through deep supervision and scaling (Mur-Labadia et al., 2026); whether scaling reduces this trade-off remains untested. Alignment during pretraining rather than a separate phase is unexplored and a candidate for future work.

Finally, our evaluation focuses on probing, while SSL representations are increasingly used in generative and multimodal systems (Appendix H.3). This motivates broader evaluation of properties such as reconstruction, their use for generation and generation alignment, and multimodal adaptability, supported by high-quality, diverse, and licensed music datasets.

## 6 Conclusion

We introduced MultiJEPA, a multiresolution self-supervised framework for learning music representations within a single encoder. A controlled fixed-resolution sweep exposes a context–detail tradeoff: coarse views favor tonal and structural understanding, while fine views favor timing and transcription. MultiJEPA broadly recovers these complementary operating points by sharing parameters across resolution levels, and parameter-free fusion extends the observed frontier. A short self-alignment phase improves cross-resolution consistency and probe transfer, although native-task performance is not uniformly preserved. These findings demonstrate that multi-resolution self-supervised learning within a single encoder can match or exceed specialized fixed-resolution models across a broad evaluation suite. Results are shown for music but our method is applicable in other domains with broad separation of semantics across different timescales.

## AI USE STATEMENT

We used OpenAI Codex, and Claude early in the project, to assist with experimental code and plotting, discuss experimental design and result interpretation, check literature and citations, and draft and proofread parts of the manuscript. The authors reviewed and verified all AI-assisted work and take full responsibility for the paper's content, including its claims and figures.

## REPRODUCIBILITY STATEMENT

Appendix A describes the training settings and encoder architecture. Appendix B documents the downstream probing protocol, with task-specific details in Appendices B.2 and B.3. Appendices F and G describe the cross-resolution alignment and context-generalization evaluations, and Appendix H defines the intrinsic metrics. We cannot release the training data or weights, but provide details to support reproduction with comparable data and compute.

## REFERENCES


Taketo Akama, Hiroaki Kitano, Katsuhiro Takematsu, Yasushi Miyajima, and Natalia Polouliakh. Auxiliary self-supervision to metric learning for music similarity-based retrieval and auto-tagging. *PLOS ONE*, 18(11):e0294643, 2023. doi: 10.1371/journal.pone.0294643.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Komeili Mojtaba, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024.

Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14496–14506. IEEE, 2023.

Rachel M. Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan*, pp. 155–160, 2014.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Rodrigo Castellon, Chris Donahue, and Percy Liang. Codified audio language modeling learns useful representations for music information retrieval. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, pp. 88–96, 2021. URL https://archives.ismir.net/ismir2021/paper/000010.pdf.

Chun-Fu (Richard) Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 357–366, 2021.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. URL https://arxiv.org/abs/2407.10759.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024.

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 (ISMIR)*, pp. 316–323, Suzhou, China, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017. doi: 10.1038/s41598-017-11873-y.

Zhengcong Fei, Mingyuan Fan, and Junshi Huang. A-jepa: Joint-embedding predictive architecture can listen. *arXiv preprint arXiv:2311.15830*, 2023.

Jiu Feng, Mehmet Hamza Erol, Joon Son Chung, and Arda Senocak. FlexiAST: Flexibility is what AST needs. In *Proceedings of Interspeech 2023*, pp. 2828–2832, 2023. doi: 10.21437/Interspeech.2023-745. URL https://www.isca-archive.org/interspeech_2023/feng23c_interspeech.html.

Jiu Feng, Mehmet Hamza Erol, Joon Son Chung, and Arda Senocak. ElasticAST: An audio spectrogram transformer for all length and resolutions. In *Interspeech 2024*, pp. 4743–4747, 2024. doi: 10.21437/Interspeech.2024-1890. URL https://www.isca-archive.org/interspeech_2024/feng24c_interspeech.html.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 2023.

Yicheng Gu, Junan Zhang, Jerry Li, Zhizheng Wu, and Lauri Juvela. Frequency-aware self-supervised music representation learning. *arXiv preprint arXiv:2606.25713*, 2026.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Eric J. Humphrey, Simon Durand, and Brian McFee. OpenMIC-2018: An open data-set for multiple instrument recognition. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 438–444, Paris, France, 2018.

Angelos-Nikolaos Kanatas, Yuexuan Kong, Pablo Alonso-Jiménez, Xavier Serra, and Dmitry Bogdanov. What makes a good layer? assessing the layer-wise intrinsic properties of music foundation models. *arXiv preprint arXiv:2608.14819*, 2026.

Corey Kereliuk, Bob L. Sturm, and Jan Larsen. Deep learning and music adversaries. *IEEE Transactions on Multimedia*, 17(11):2059–2071, 2015. doi: 10.1109/TMM.2015.2478068.

Peter Knees, Ángel Faraldo, Perfecto Herrera, Richard Vogl, Sebastian Böck, Florian Hörschläger, and Mickael Le Goff. Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 364–370, Málaga, Spain, 2015.

Yuexuan Kong, Vincent Lostanlen, Gabriel Meseguer-Brocal, Stella Wong, Mathieu Lagrange, and Romain Hennequin. Stone: Self-supervised tonality estimator. *Proceedings of International Society for Music Information Retrieval Conference (ISMIR 2024)*, 2024.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

Edith Law, Kris West, Michael I. Mandel, Mert Bay, and J. Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 387–392, Kobe, Japan, 2009.

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 (ICLR)*, 2024.

Bin Lin, Bo Zhao, Boyang Wang, Boyang Zhang, Boyong Wu, Chao Yan, Chen Geng, Chen Wu, Cheng Yi, Chengli Feng, et al. Stepaudio 3 gen technical report. *arXiv preprint arXiv:2609.12945*, 2026.

Siyu Liu, Chujie Qin, Hubery Yin, Qixin Yan, Zheng-Peng Duan, Chen Li, Jing Lyu, Chun-Le Guo, and Chongyi Li. Improving reconstruction of representation autoencoder. *arXiv preprint arXiv:2602.08620*, 2026.

Ben Maman and Amit H. Bermano. Unaligned supervision for automatic music transcription in the wild. In *Proceedings of the 39th International Conference on Machine Learning, ICML 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14918–14934. PMLR, 2022.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Oriol Nieto, Matthew C. McCallum, Matthew E. P. Davies, Andrew Robertson, Adam M. Stark, and Eran Egozy. The harmonix set: Beats, downbeats, and functional segment annotations of western popular music. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands*, pp. 565–572, 2019.

Mang Ning, Mingxiao Li, Le Zhang, Lanmiao Liu, Matthew B. Blaschko, Albert Ali Salah, and Itir Onal Ertugrul. Spectrum matching: a unified perspective for superior diffusability in latent diffusion. *arXiv preprint arXiv:2603.14645*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

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.

Aurian Quelennec, Pierre Chouteau, Geoffroy Peeters, and Slim Essid. Masked latent prediction and classification for self-supervised audio representation learning. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025a. doi: 10.1109/ICASSP49660.2025.10887666.

Aurian Quelennec, Pierre Chouteau, Geoffroy Peeters, and Slim Essid. MATPAC++: Enhanced masked latent prediction for self-supervised audio representation learning. *arXiv preprint arXiv:2508.12709*, 2025b.

Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis. mir_eval: A transparent implementation of common MIR metrics. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 367–372, Taipei, Taiwan, 2014.

Simon Rouard, Yossi Adi, Jade Copet, Axel Roebel, and Alexandre Défossez. Audio conditioning for music generation via discrete bottleneck features. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 146–153, 2024.

Jiatong Shi, Hirofumi Inaguma, Xutai Ma, Ilia Kulikov, and Anna Sun. Multi-resolution HuBERT: Multi-resolution speech self-supervised learning with masked unit prediction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6d09ef61aeb76be676b358f6f87b3484-Abstract-Conference.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *International Conference on Learning Representations (ICLR)*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3929a7785bd56f57edcff0152ab41289-Abstract-Conference.html.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026b.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *International Conference on Machine Learning (ICML)*, 2025.

Bob L. Sturm. The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

John Thickstun, Zaïd Harchaoui, and Sham M. Kakade. Learning features of music from scratch. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=rkFBJv9gg.

Ludovic Tuncay, Etienne Labbé, Emmanouil Benetos, and Thomas Pellegrini. Audio-JEPA: Joint-embedding predictive architecture for audio representation learning. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2025.

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.

Marcel A. Vélez Vásquez and John Ashley Burgoyne. Tailed U-Net: Multi-scale music representation learning. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference*, pp. 67–75, 2022. doi: 10.5281/zenodo.7316596. URL https://archives.ismir.net/ismir2022/paper/000007.pdf.

Minz Won, Yun-Ning Hung, and Duc Le. A foundation model for music informatics. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1226–1230. IEEE, 2024.

Qingyang Xi, Rachel M. Bittner, Johan Pauwels, Xuzhou Ye, and Juan Pablo Bello. Guitarset: A dataset for guitar transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France*, pp. 453–460, 2018.

Jin Xu, Kangdi Wang, Ruibin Yuan, Shun Lei, Xiong Wang, Xize Cheng, Xueyao Zhang, Yang Zhang, Yiheng Chen, Yongqi Wang, et al. Qwen-music technical report. *arXiv preprint arXiv:2607.11699*, 2026.

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 B. 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 36 (NeurIPS 2023) Datasets and Benchmarks Track*, 2023.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations (ICLR)*, 2026.

Tianxiong Zhong, Xingye Tian, Xuebo Wang, Xin Tao, and Pengfei Wan. Diffusing in the right space: A systematic study of latent diffusability. *arXiv preprint arXiv:2606.03578*, 2026.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. MuQ: Self-supervised music representation learning with mel residual vector quantization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2025.

# A Architecture and Training

The fixed-resolution and multiresolution models share a masked feature-prediction objective. Their main architectural difference is the frontend: the multiresolution model resamples a common fine patch grid before the encoder.

## A.1 JEPA encoder and deep supervision

Stereo audio is represented at 44.1 kHz by a two-channel, 128-bin log-mel tensor with a 441-sample hop, a 2048-sample Fourier transform, and frequencies from 50 Hz to 14 kHz. The finest patchifier divides the spectrogram into non-overlapping $4 \times 16$ time–frequency patches. A linear projection maps each $2 \times 4 \times 16$ patch to an embedding token.

Let $M$ be the masked token indices and $C$ their complement. The student encoder $E_\theta$ processes only $z_C$, whereas the teacher $E_{\bar{\theta}}$ processes the complete sequence $z$. Its parameters follow the exponential moving average

$$\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta. \tag{3}$$

For each example, we use either random masking or contiguous block masking with equal probability and draw the masking ratio uniformly from $[0.7, 0.9]$.

A lightweight predictor $P_\phi$ receives the student features and learned mask tokens $\Delta$ inserted at $M$. The student prediction and teacher target sequences are

$$\hat{y} = P_\phi(E_\theta(z_C), \Delta), \qquad \bar{y} = E_{\bar{\theta}}(z).$$

For an encoder with $L$ layers, deep supervision uses

$$\mathcal{D} = \{L/4, L/2, 3L/4, L\}, \qquad \mathcal{L}_{\text{deep}} = \frac{1}{|\mathcal{D}|} \sum_{\ell \in \mathcal{D}} \mathcal{L}^{(\ell)}_{\text{mask}}. \tag{4}$$

The student states at these depths are concatenated channel-wise and projected once into the predictor dimension. One shared predictor processes the fused sequence, and separate output projections recover the teacher target at each depth. The four losses are weighted equally.

Table 3: Semantic encoder configurations. Parameter counts include the encoder frontend and, for multiresolution models, the TRB; they exclude the predictor and EMA copy.

| Encoder | Layers | Width | FF width | Heads | Parameters |
|---|---|---|---|---|---|
| Multi-res. Small | 12 | 384 | 1536 | 6 | ∼31M |
| Multi-res. Base | 12 | 768 | 3072 | 12 | ∼92M |
| Fixed-res. Large | 24 | 1024 | 4096 | 16 | ∼403M |
| Multi-res. Large | 24 | 1024 | 4096 | 16 | ∼420M |

Both encoder and predictor use four register tokens. The predictor follows the small-predictor design of V-JEPA (Bardes et al., 2024): six transformer layers of width 384, six attention heads, and a feed-forward width of 1536.

## A.2 Multiresolution patchification

The Transformer Resampling Block (TRB) turns each local group of fine patches into one token. Changing the group size changes the temporal rate while keeping the same backbone and embedding dimension.

Let $X_g \in \mathbb{R}^{a_r \times D}$ denote one temporal group of finest-grid tokens at a fixed frequency position, where $a_r \in \{1, 2, 4, 8\}$ and $D = 1024$ for the Large model. For $a_r > 1$, the TRB appends a shared learned query $q$ and applies one pre-normalized transformer block independently to each group:

$$\begin{aligned} H_g^{(0)} &= [X_g; q], \\ H'_g &= H_g^{(0)} + \mathrm{MHA}\Big(\mathrm{RMSNorm}(H_g^{(0)})\Big), \\ H_g^{(1)} &= H'_g + \mathrm{SwiGLU}\big(\mathrm{RMSNorm}(H'_g)\big). \end{aligned} \tag{5}$$

Only the state corresponding to $q$ is retained. This is block-local self-attention over the fine tokens and learned query, rather than cross-attention, convolution, or fixed pooling. The TRB has width 1024, eight attention heads, a 4096-dimensional feed-forward layer, RMS-normalized queries and keys, and no dropout. It does not use positional embeddings internally; locality is imposed by the block attention pattern. We observed comparable performance with and without RoPE in the TRB. The semantic transformer instead receives continuous two-dimensional RoPE coordinates, whose temporal increments are scaled by $a_r$. The finest query bypasses the TRB. multiresolution pretraining fixes the backbone sequence length across rates, while patchification and resampling costs remain rate-dependent.

The patch projection, TRB, and semantic transformer are shared across rates; no rate-specific parameters are introduced. The student and EMA teacher contain separate copies of the complete encoder, and the TRB parameters participate in the EMA update. This learned resampling should also be distinguished from the fixed block mean $\rho_{f\to c}$ in the alignment objective: the latter acts on post-encoder features and has no trainable parameters.

Table 4: Rate-dependent tensor geometry. Encoder pretraining keeps 2048 output tokens while varying crop duration. The nominal 3.1-Hz rate is exactly 3.125 Hz.

| Rate | $a_r$ | Pretraining crop | Finest-grid patches | Encoder output |
|---|---|---|---|---|
| 25 Hz | 1 | 10.24 s | $256 \times 8$ | 2048 |
| 12.5 Hz | 2 | 20.48 s | $512 \times 8$ | 2048 |
| 6.25 Hz | 4 | 40.96 s | $1024 \times 8$ | 2048 |
| 3.125 Hz | 8 | 81.92 s | $2048 \times 8$ | 2048 |

### A.3 Training configuration

Table 5 lists the full hyperparameter configuration for pre-training and the alignment phase. Self-alignment and the continued-pretraining control each receive 30,000 additional updates from the same 400,000-step checkpoint.

Table 5: Training configuration. Pre-training values apply to all model scales (S/B/L); alignment values describe multiresolution self-alignment. The 50-Hz condition belongs only to the fixed-resolution sweep.

| Optimization | Pre-training | Alignment |
|---|---|---|
| Optimizer | AdamW | |
| $(\beta_1, \beta_2)$ | $(0.9, 0.95)$ | |
| $\epsilon$ | $10^{-8}$ | |
| Weight decay | 0.04 | |
| LR schedule | Cosine decay | |
| Schedule peak LR | $5\times10^{-4}$ | |
| Minimum LR | $10^{-6}$ | |
| Warmup steps | 50,000 | None (resumed) |
| Total steps | 400,000 | 30,000 |
| Gradient clipping | $\|\nabla\|_2 \le 5.0$ | |
| Effective batch | 1,024 | |
| GPUs | 32 | |

| Encoder and losses | Pre-training | Alignment |
|---|---|---|
| EMA decay $\tau$ | .99995→.99999 | .99999 |
| EMA warmup | 100,000 steps | — |
| Mask strategy | Random or block ($p = .5$) | |
| Masking ratio | $\mathcal{U}[.7, .9]$ | |
| Grid tokens | 2,048 | 2,048–16,384 |
| Crop duration | Rate-dependent | teacher |
| Resolution set | $\{4, 8, 16, 32\} \times 16$ | |
| DeepSSL depths | $\{L/4, L/2, 3L/4, L\}$ | |
| Register tokens | 4 | |
| $\lambda_{\text{feat}}$ | — | 1.0 |
| Context supervision | ✗ | |

**Self-alignment crop and token geometry.** During self-alignment, both views use the crop length of the teacher resolution. Block pooling maps the fine output onto the coarse grid over the same audio. Alongside alignment, this phase therefore trains the finer queries on longer contexts and more tokens than they see during pretraining.

The JEPA objective uses masked student inputs. Feature alignment uses a separate, unmasked student forward pass, whose fine-grid features are block-pooled and matched to the unmasked coarse EMA-teacher features. We accumulate gradients from both losses before a single optimizer update.

## B Datasets and Evaluation Protocols

The evaluation suite tests both sides of the context–detail tradeoff: track-level semantics and temporally precise musical events. Table 6 lists the datasets and their metrics; all evaluations keep the encoder frozen.

Table 6: Downstream evaluation suite. We distinguish **fine** tasks (bottom) and **coarse** tasks (top).

| Dataset | Task and level | Primary metric |
|---|---|---|
| GTZAN | Genre classification (clip) | Accuracy |
| GiantSteps | Key estimation (song) | MIREX weighted |
| FMAKv2 | Key estimation (clip) | MIREX weighted |
| MagnaTagATune | Auto-tagging (clip) | AUROC / mAP |
| OpenMIC | Instrument detection (clip) | mAP |
| Harmonix | Structural segmentation (song) | mIoU, pairwise F, boundary F |
| GuitarSet | Note transcription (frame) | Frame F1, Onset F1 |
| GTZAN-Beats | Beat tracking (frame) | F1 @ 70 / 20 ms |
| MusicNet-EM | Note transcription (frame) | Frame F1, Onset F1 |
| MedleyDB F0 | Predominant / multi-pitch F0 (frame) | RPA / multipitch F1 |

**Representation extraction.** Each dataset–representation pair receives an independent linear probe trained with Adam and selected by its validation metric. Frame probes apply the same head independently at every frame, without temporal layers. Clip embeddings average deterministic windows over the full track; frame embeddings stitch windows chronologically and resample to the annotation grid. multiresolution fusion uses fixed equal-weight averaging after interpolation to the finest grid.

We select the layer with the highest estimated ID, yielding layer 20 of 24 for our Large encoder. The selected layer is shared across downstream tasks, with the same selection rule applied to external encoders. This is a task-independent rule, not an estimate of the optimal layer for every task, and is separate from validation-based early stopping of the downstream head. Clip-level affine probes use a learning rate of $5 \times 10^{-3}$, weight decay $10^{-4}$, and batch size 64, for at most 100 epochs. Training stops after 20 epochs without validation improvement and restores the best-validation probe weights. Reported means average five probe runs, not five independently pretrained encoders.

### B.1 Clip-level and beat protocols

**GTZAN.** Genre classification uses the fault-filtered split and reports top-1 accuracy (Tzanetakis & Cook, 2002; Sturm, 2013; Kereliuk et al., 2015).

**GiantSteps and FMAKv2 key.** GiantSteps (Knees et al., 2015) averages encoder-window embeddings within each dataset clip, then averages probe logits across clips from the same song for song-level scoring. FMAKv2 (Kong et al., 2024; Defferrard et al., 2017) uses a deterministic split. Both report the MIREX weighted key score (Raffel et al., 2014; Yuan et al., 2023).

**MagnaTagATune and OpenMIC.** Tagging and instrument detection use macro AUROC and average precision after removing unlabeled examples (Law et al., 2009; Humphrey et al., 2018).

**GTZAN-Beats.** A linear frame probe feeds MARBLE-compatible DBN decoding; F-measure is reported at 70 and 20 ms tolerances (Tzanetakis & Cook, 2002; Yuan et al., 2023).

### B.2 Structural segmentation

Structural segmentation tests whether features recover repeated musical sections and their boundaries across a track. The results below distinguish organization already present in the features from performance after learning a task-specific projection or head.

Section annotations are reduced to seven functional classes. We use three Harmonix (Nieto et al., 2019) structure protocols referenced in Section 3.2. All operate on frame representations at the Harmonix 2 Hz label grid and macro-average metrics over test tracks.

**Joint linear probe** One linear head predicts seven structural functions and a boundary logit using cross-entropy plus class-balanced binary cross-entropy. Peak-picked boundaries define segments whose function logits are mean-pooled. We report boundary F at 0.5 and 3 s, pairwise F, and frame accuracy.

**Zero-shot segmentation (oracle-$K$, zero learned probe)** We apply oracle-$K$ k-means directly to frozen frame representations, smooth the cluster sequence, merge short runs, and map clusters to reference functions by Hungarian assignment (Kuhn, 1955). Tables report the post-processed mIoU and pairwise F. Labels supply $K$ and the final assignment, so this is zero-probe rather than fully label-free.

**Relational probe** The relational probe learns a low-rank affinity-preserving projection:

$$z_t = \text{normalize}(Wx_t), \qquad s_{ij} = \cos(z_i, z_j)/\tau + b,$$

where $W$ is a bias-free 64-dimensional projection, $\tau$ a learned temperature, and $b$ a learned bias. BCE on balanced same/different-function pairs trains the projection. We report recurrence ranking from the learned affinities and segmentation after applying the zero-shot clustering pipeline to $z_t$.

### B.3 Transcription and melody

Frame transcription measures which pitches are active over time, while onset scoring tests note-start localization. Reporting both distinguishes retained pitch information from timing precision.

**GuitarSet and MusicNet-EM.** GuitarSet (solo guitar, clean polyphonic transcription) and MusicNet-EM (classical ensembles, expectation-maximization labels) form the composite reported in Table 2 (Xi et al., 2018; Thickstun et al., 2017; Maman & Bermano, 2022). Each is evaluated with frame F1, chroma F1, onset F1, onset+offset F1, and YourMT3 frame F1 under the standard mir_eval note protocol at 16 ms resolution. The transcription composite averages frame and onset F1 across the two datasets.

**MedleyDB F0.** Melody2 raw pitch accuracy measures predominant F0; Multi-F0 F1 measures annotated melodic-source activity (Bittner et al., 2014).

## C Qualitative Visualizations

These visualizations complement the probe scores by showing where representations change within a track and where sections recur.

We project features onto three principal components and map them to RGB timelines. In Figure 8, the 3.1-Hz view follows repeated sections, while finer views contain denser variation within them. Similarity to a chorus anchor recovers track-scale recurrence; the linked zooms compare local kick and vocal similarity at coarse and fine rates.

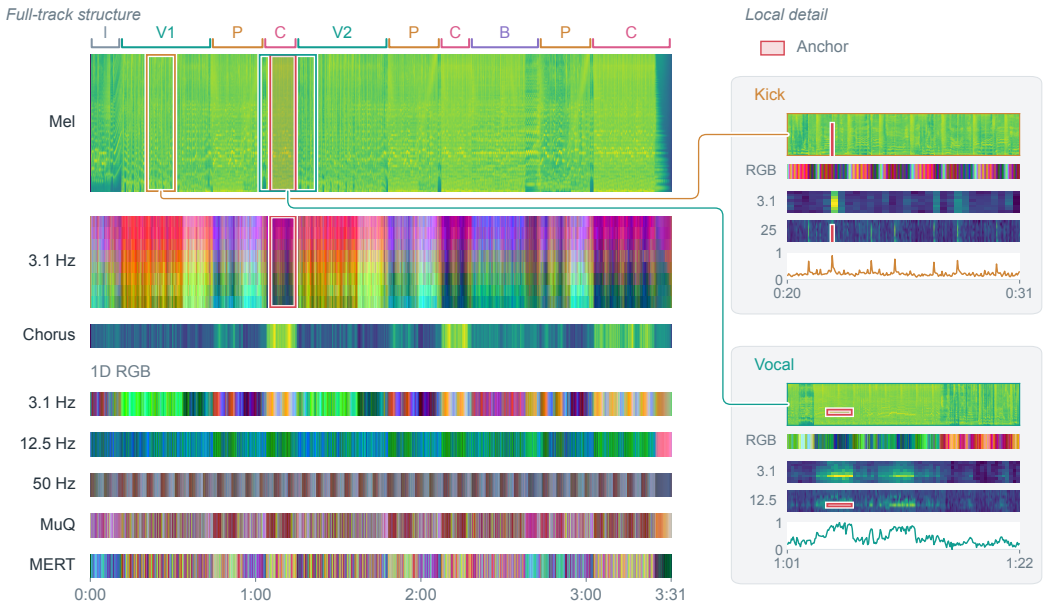


Figure 8: Qualitative visualization of Structure and local detail in a rock track. *Left:* full-track spectrogram, 3.1-Hz PCA-RGB features, chorus similarity, and RGB timelines comparing three JEPA rates with MuQ and MERT. Section brackets mark intro (I), verse (V), pre-chorus (P), chorus (C), and bridge (B). *Right:* linked kick and vocal zooms compare coarse (3.1 Hz) and fine (25/12.5 Hz) features. Red boxes select anchors; maps show cosine similarity to their mean embedding, and profiles average fine-grid similarity over the anchor frequency band.

Figure 9 adds three manually selected passages from *Locked Out of Heaven* (Bruno Mars). The middle panel uses a rimshot anchor. The profiles measure feature similarity to each selected region; they are not supervised event predictions.

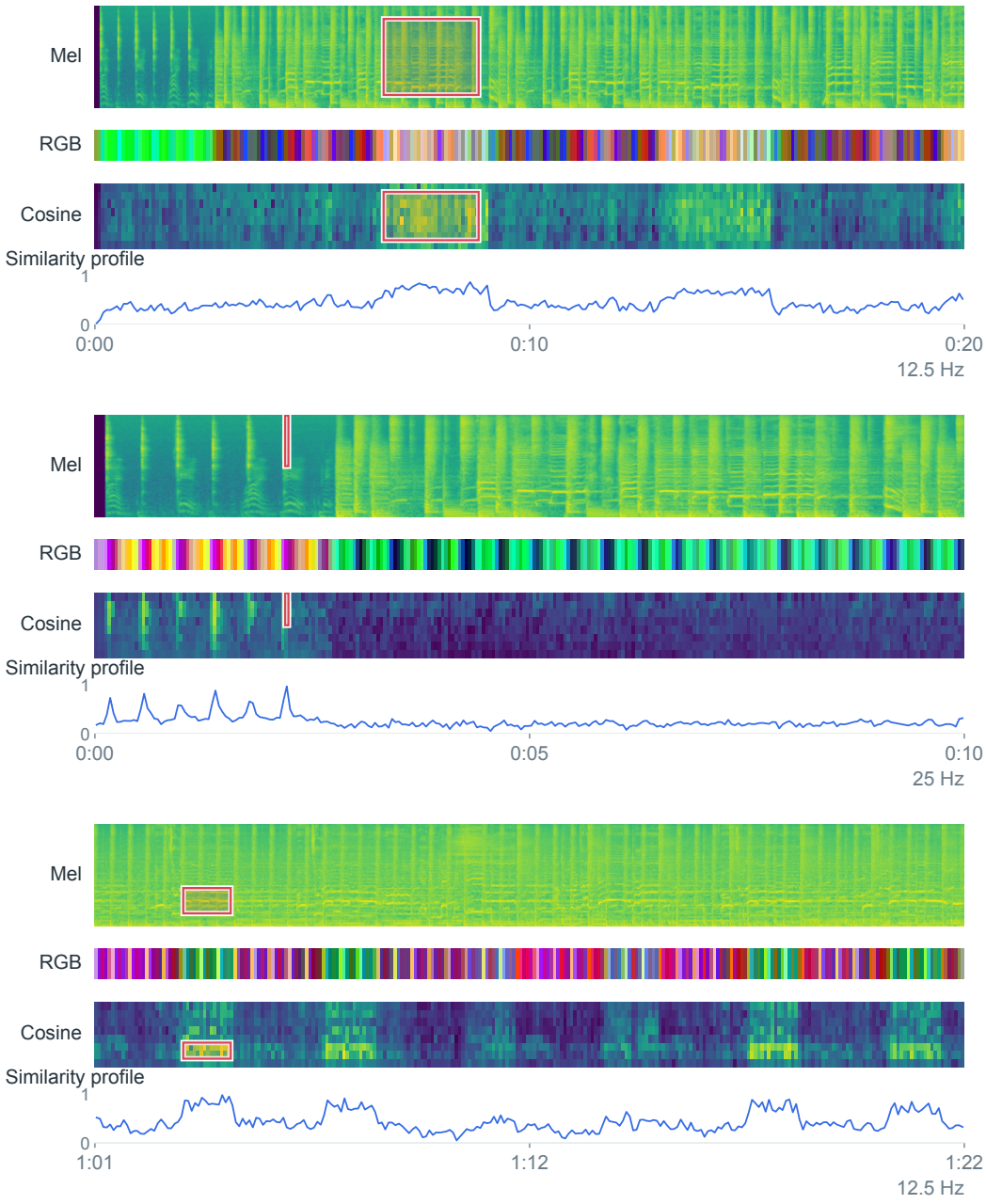


Figure 9: Additional qualitative excerpts from *Locked Out of Heaven*. Each panel shows the mel spectrogram, a one-dimensional PCA-RGB timeline, cosine similarity to the outlined anchor, and a similarity profile averaged over the selected frequency band. RGB timelines average raw frequency embeddings before fitting PCA within each crop.

## D Ablations and Scaling

**Visible-token context supervision.** The main JEPA loss supervises masked positions. We test an auxiliary loss on visible tokens to reduce the outlier artifacts illustrated in Figure 10; this loss is not used in the main recipe. Table 7 compares it with changes to supervision depth and crop duration in a fixed-resolution model.

Context supervision adds a loss on visible positions $C$, using the same predictor and stop-gradient EMA-teacher targets at the four supervised depths $\mathcal{D}$. The student still receives only visible tokens, while the teacher processes the complete grid. For integer time–frequency grid coordinates $p_i$, let $\delta_i = \min_{j \in M} \|p_i - p_j\|_2$ be the distance to the nearest masked position. With feature dimension $d$, the added loss is

$$\mathcal{L}_{\text{ctx}} = \frac{\lambda_{\text{ctx}}}{|\mathcal{D}|\,|C|\,d} \sum_{\ell \in \mathcal{D}} \sum_{i \in C} \frac{\|\hat{y}_i^{(\ell)} - \text{sg}(\bar{y}_i^{(\ell)})\|_2^2}{\sqrt{\max(1, \delta_i)}}, \qquad \mathcal{L} = \mathcal{L}_{\text{deep}} + \mathcal{L}_{\text{ctx}}. \tag{6}$$

Context supervision starts at 300k steps; its weight rises linearly from 0 to 0.5 over 50k updates, then stays constant.

Table 7: Matched $32 \times 16$ ablations. Each model uses its single ID-selected layer.

| Configuration | GTZAN | Key | $\text{Beat}_{70}$ | $\text{Beat}_{20}$ | mIoU |
|---|---|---|---|---|---|
| JEPA + DeepSSL | 84.8 | 64.3 | 40.0 | 15.6 | 49.0 |
| − DeepSSL | 77.6 | 56.0 | 35.3 | 12.9 | 45.7 |
| Shorter context | 80.0 | 53.3 | 36.3 | 13.6 | 42.2 |
| + Context supervision | 81.7 | 65.0 | 36.4 | 14.3 | 47.6 |

Deep supervision improves semantic and rhythmic performance. Visible-token context supervision instead changes the task balance: key improves, while genre, structure, and beat scores decline.

### D.1 Token-level artifacts

Self-supervised ViTs are known to develop high-norm outlier tokens that repurpose low-information patches as global-information storage (Darcet et al., 2024). We observe analogous artifacts in MultiJEPA: PCA visualizations show outlier tokens emerging after ∼200k training steps, concentrated in low-energy mel frequency bands and more visible under chunked inference than full-track single-pass inference.

Supervising visible tokens reduces the artifacts during chunked inference, with a small gain on key but lower genre, structure, and beat scores (Table 7).

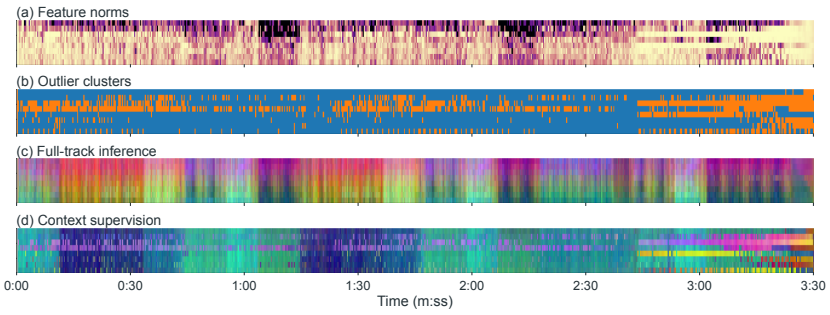


Figure 10: Token-level artifacts and two mitigations, on a full pop track. Top to bottom: per-patch feature norms; an outlier-cluster map; PCA-RGB features under one full-attention pass, which reduces the chunking artifact; and the context-supervision variant, which reduces it with a modest task tradeoff (Table 7).

### D.2 Model scaling

Small and Base use the same multiresolution recipe at lower capacity. Tables 8 and 9 show how much performance these smaller encoders retain, distinguishing the best individual query for each task from four-query fusion.

Table 8: Model scaling with data, training recipe, rates, and token count held fixed. Scores use the best native query.

| Size | Params. | GTZAN | FMAKv2 | Key | $F1_{70}$ | $F1_{20}$ |
|---|---|---|---|---|---|---|
| Small | 31M | 76.9 | **66.5** | **61.8** | 87.4 | 65.1 |
| Base | 92M | 76.6 | 66.1 | 61.1 | 86.1 | 64.5 |
| Large | 420M | **88.3** | 65.4 | 61.7 | **90.4** | **68.8** |

At this fixed budget, Large improves genre and beat substantially but not key; Small and Base remain close. Capacity therefore changes some task families more than others rather than shifting all operating points uniformly.

Table 9 reports the four-query fusion scores underlying the Small and Base points in Figure 1, whereas Table 8 reports the best native query for each task.

Table 9: Small and Base four-query fusion: raw Pareto constituents, in percent. Transcription scores average GuitarSet and MusicNet-EM; F0 uses MedleyDB Melody2 RPA and Multi-F0 F1. Bold indicates the higher score in each column.

| | Context and structure | | | Beat | | Transcription | | F0 | |
|---|---|---|---|---|---|---|---|---|---|
| Size | GTZAN Acc. | Key MIREX | Harmonix Rel. mIoU | $F1_{70}$ | $F1_{20}$ | Frame F1 | Onset F1 | RPA | Multi-F0 F1 |
| Small | 76.6 | **61.2** | 40.9 | 87.0 | 64.6 | 74.1 | **56.8** | 58.2 | 46.7 |
| Base | **77.6** | 60.8 | **42.6** | **87.2** | **66.3** | **75.1** | 56.3 | **59.4** | **47.0** |

## E Sensitivity to Task Coverage

Figure 11 tests the sensitivity of the frontier in Figure 1, using all sixteen headline metrics in Table 2. The six local-detail terms are unchanged. The context–structure composite adds FMAKv2 MIREX, OpenMIC mAP, Harmonix relational pairwise F, and zero-shot Harmonix mIoU and pairwise F. Including zero-shot structure gives weight to organization already present in the embeddings, without a learned relational projection.

For model $i$ and metric $m$, we normalize its score $s_{im}$ using a fixed reference cohort $\mathcal{C}_m$ and average over the metrics $\mathcal{M}$ on each axis:

$$\tilde{s}_{im} = \frac{s_{im} - \min_{j \in \mathcal{C}_m} s_{jm}}{\max_{j \in \mathcal{C}_m} s_{jm} - \min_{j \in \mathcal{C}_m} s_{jm}}, \qquad S_i(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \tilde{s}_{im}.$$

The nine original metrics use the five fixed-resolution models, four native queries, Large/Small/Base four-query fusion, and seven external baselines (MERT, MuQ, MusicFM, MATPAC++, PupuJEPA, SAME-L, and SAO) to set the bounds. These bounds remain fixed. The seven added metrics use the same cohort without Small and Base, whose saved comparison lacks these columns. MERT-v2 and self-aligned fusion do not change the bounds, and values outside the reference range are not clipped. This comparison broadens task coverage but also changes the weighting: Harmonix contributes four of ten context–structure terms and MagnaTagATune two. It is therefore a sensitivity analysis of the composite, not a uniquely defined task balance. Small and Base are omitted because their saved comparison lacks the additional columns.

Four-query fusion remains nondominated under this broader composite, together with the 3.1- and 6.25-Hz fixed-resolution models. Self-aligned fusion lies below the original fusion point but remains above and to the right of the external encoders. The larger separation from MERT-v2 reflects the

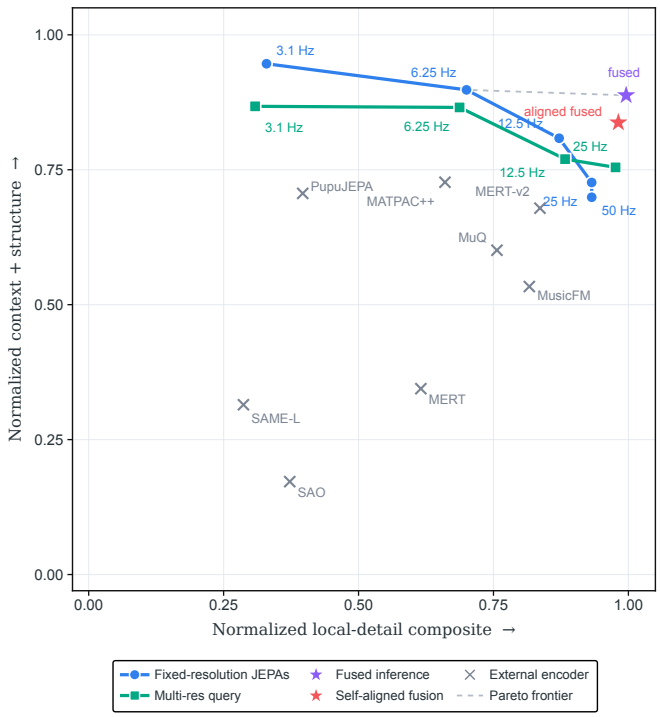


Figure 11: Alternative context–detail frontier using all sixteen headline metrics. The local-detail axis is unchanged; the context–structure axis includes tagging, instrument detection, both key datasets, and relational and zero-shot structure. Self-aligned fusion is included under the same normalization.

inclusion and weighting of the additional metrics, particularly zero-shot structure, rather than a change in the underlying model scores.

**Self-aligned queries and fusion.** Using the same task composites and fixed normalization as Figure 1, we compare self-aligned queries and fusion with their parent, using the single ID-selected layer for each model. Relative to the matched parent, the four native queries change by $-0.002$ to $+0.007$ in local detail and gain $0.004$ to $0.024$ in context and structure. Equal-weight fusion changes by $-0.001$ and $-0.012$, respectively. Self-alignment therefore improves the context composite of the individual rate-conditioned views without extending the original headline frontier; fusion remains the strongest aligned local-detail operating point.

## F Self-Alignment and Probe Transfer

Self-alignment trains block-pooled fine features to match a native coarse view of the same audio. We examine how quickly the views agree, whether their native task performance changes, and whether a coarse-view probe can be reused on pooled fine features without refitting.

**Encoder branches.** Context generalization and probe transfer evaluate the directly optimized student. CKNNA compares its pooled fine features with native coarse EMA-teacher targets. The native-probe and intrinsic trajectories instead evaluate the EMA teacher. With decay $\tau = 0.99999$, the 30k alignment updates receive total weight $1 - \tau^{30000} \approx 26\%$ in the final teacher parameters. This is an EMA parameter weight, not a fraction of the effect on representations; the student gives the direct effect of the alignment phase.

**Training dynamics.** The alignment objective shows only a short transition before decreasing over the remainder of the phase (Figure 12a).

Native geometry stays largely stationary while cross-resolution agreement rises. Over the first 20k alignment steps, across the four native resolutions and fusion, TwoNN changes by at most 0.7%, RankMe by 3.1%, and LDS by 7.7%. After 30k steps, native-query changes remain within 2.4 points on GTZAN, 1.3 points on Beat $F1_{70}$, and 0.9 points on key; fusion changes by 0.0, $-0.3$, and $+0.4$ points. Harmonix relational mIoU changes by $-0.2$ to $+2.2$ points across native queries and improves by 1.4 points for fusion. For these EMA results, alignment changes cross-resolution agreement much more than native geometry or probe scores.

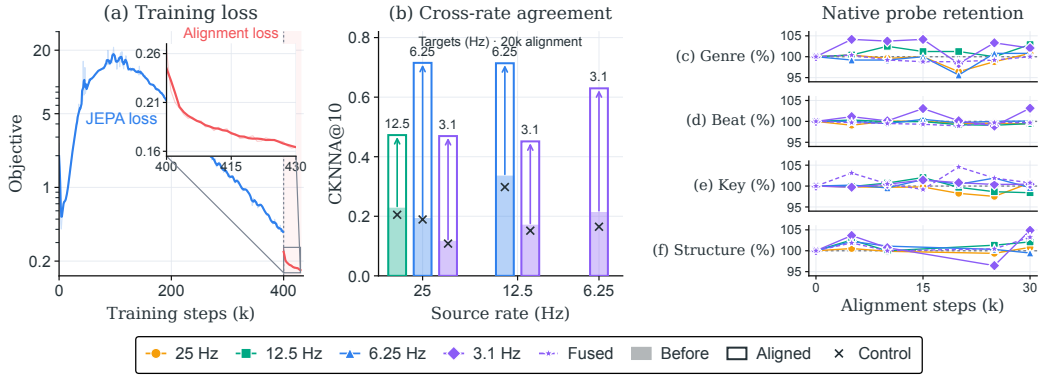


Figure 12: Evolution during self-alignment. **(a)** JEPA training loss, with the separate feature-alignment objective shown over the highlighted phase in the inset. Thin lines are 1k-step means, thick lines five-bin moving averages. **(b)** Student-to-teacher CKNNA before and after 20k alignment steps, grouped by source rate; colours and bar labels identify targets. Crosses show continued pretraining at the same step. **(c–f)** ID-selected-layer EMA native-query and fused probe scores, relative to each view's alignment-start score (100%): genre accuracy, micro Beat $F1_{70}$, song-level key MIREX, and Harmonix relational mIoU.

### F.1 Complete probe-transfer trajectories

Probe transfer tests whether pooling preserves the feature directions used by a native-coarse probe. At each checkpoint and for each fine-to-coarse pair, we fit a head on native-coarse student features, select it on validation performance, and reuse it unchanged on pooled-fine student features. Heads are fitted separately at each checkpoint. Retention is pooled-fine performance divided by native-coarse performance, expressed as a percentage; 100% means that pooling matches the native-coarse score. Figure 13 gives all seven task trajectories, each averaged over the six pairs; Figure 7 summarizes their endpoints.

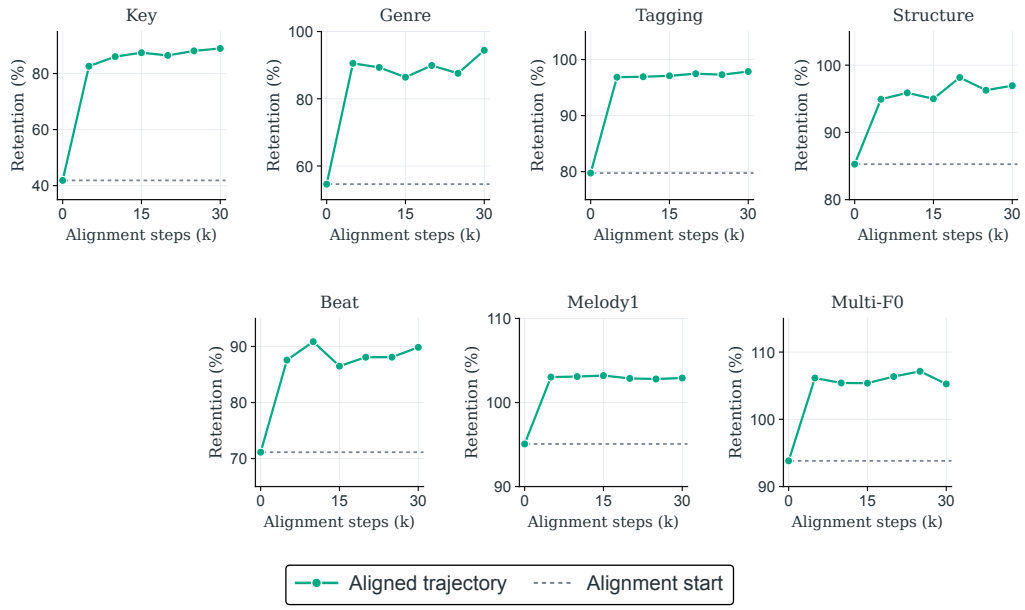


Figure 13: Complete cross-resolution probe-transfer trajectories. Each task averages all six pooled-fine/native-coarse ratios. A head is fitted separately for each pair and checkpoint, then reused unchanged across the two views.

### F.2 Neighborhood agreement

Centered Kernel Nearest-Neighbor Alignment (CKNNA) measures whether paired views preserve the same cross-track neighborhoods, without fitting a task-specific probe. Higher scores indicate stronger agreement between the views.

We use cosine-kernel CKNNA from (Huh et al., 2024), with the unbiased Hilbert–Schmidt independence criterion (HSIC) expression used in their released code. Each observation pairs two representations of the same validation track. For self-alignment, these are the final-layer student features block-pooled to the coarse grid and the native coarse features from the same checkpoint's EMA teacher. We average valid time–frequency tokens across extraction windows to obtain one vector per track and view, then apply $\ell_2$ normalization. Neighborhoods are therefore defined across tracks, not individual tokens.

For $n$ paired tracks, let $K_{ij} = x_i^\top x_j$ and $L_{ij} = y_i^\top y_j$ be the cosine kernels. Let $M^x_{ij} = \mathbb{1}[j \in \mathcal{N}^x_k(i)]$ indicate the $k$ largest non-self similarities in row $i$ of $K$, and define $M^y$ analogously. Their elementwise intersection $S = M^x \odot M^y$ retains directed neighbor relations shared by both views. For zero-diagonal matrices $A, B$, the implementation computes

$$h(A,B) = \frac{1}{n(n-3)}\left[\operatorname{tr}(AB) + \frac{(\mathbf{1}^\top A\mathbf{1})(\mathbf{1}^\top B\mathbf{1})}{(n-1)(n-2)} - \frac{2\mathbf{1}^\top AB\mathbf{1}}{n-2}\right], \tag{7}$$

where $\mathbf{1}$ is the all-ones vector. Writing $A_x = M^x \odot K$ and $A_y = M^y \odot L$, we report

$$\mathrm{CKNNA}_k(X, Y) = \frac{h(S \odot K, S \odot L)}{\sqrt{h(A_x, A_x)\, h(A_y, A_y)} + 10^{-6}}. \tag{8}$$

The masks are not symmetrized: the numerator uses their intersection, while each normalization term uses its own view's mask. Centering is handled by $h$, without subtracting feature means before cosine similarity. We use $k = 10$ for the alignment results, evaluate the HSIC reductions in float64, and report raw scores without clipping or null calibration. Mean self-alignment curves average the six resolution-pair scores with equal weight.

Figure 14 separates the six fine-to-coarse pairs to show the differences hidden by the mean trajectory in Figure 6.

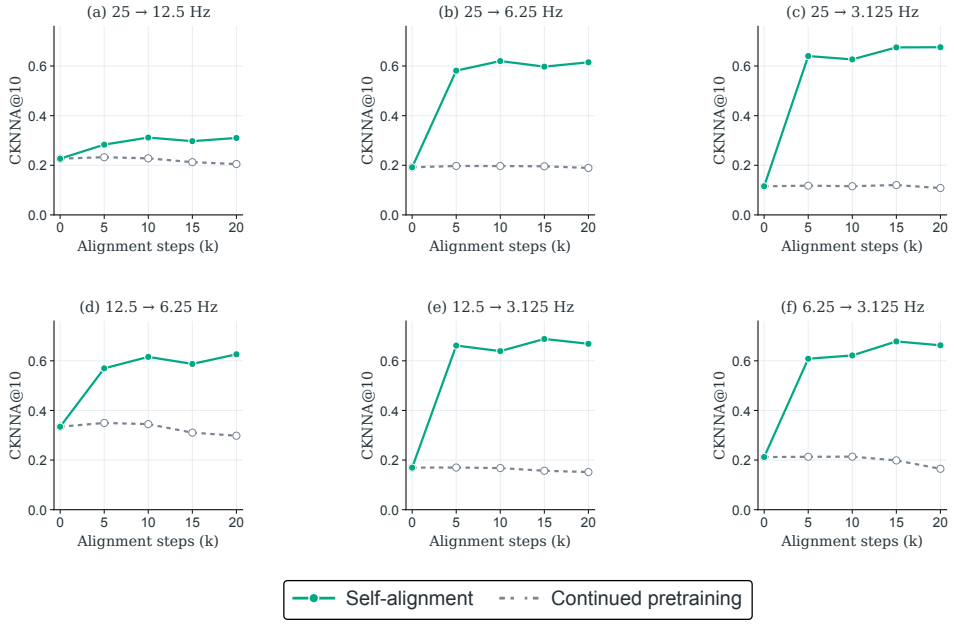


Figure 14: Pair-resolved CKNNA trajectories during self-alignment. All six fine-to-coarse pairs use the same protocol.

### F.3 Temporal spectra and diffusability

Studies of image and video latents associate stronger low-frequency structure with easier diffusion modeling (Skorokhodov et al., 2025). Spectral balance matters, however: excessive smoothing can remove useful detail (Ning et al., 2026). We use temporal power spectral density (PSD) to examine whether self-alignment strengthens slowly varying components while preserving faster variation in the fine representation.

Here frequency measures cycles per second of embedding variation, rather than waveform pitch. We average tokens across the frequency axis of the grid at each time step, estimate a one-sided Welch PSD for each embedding channel, and average the channel spectra. The estimator uses Hann-windowed segments with 50% overlap and removes each segment's mean. Within each track, extraction-window spectra are weighted by their valid temporal-frame counts; tracks then receive equal weight. The curves retain raw power per Hz, without unit-area normalization. The residual is $Y_f - \mathrm{broadcast}(\rho(Y_f))$: the fine features minus their block-pooled features repeated onto the fine grid.

Figure 15 shows stronger low-frequency power in the aligned fine representation, while its residual retains faster variation. This is consistent with the spectral motivation above; whether it improves music generation requires a trained-generator evaluation.

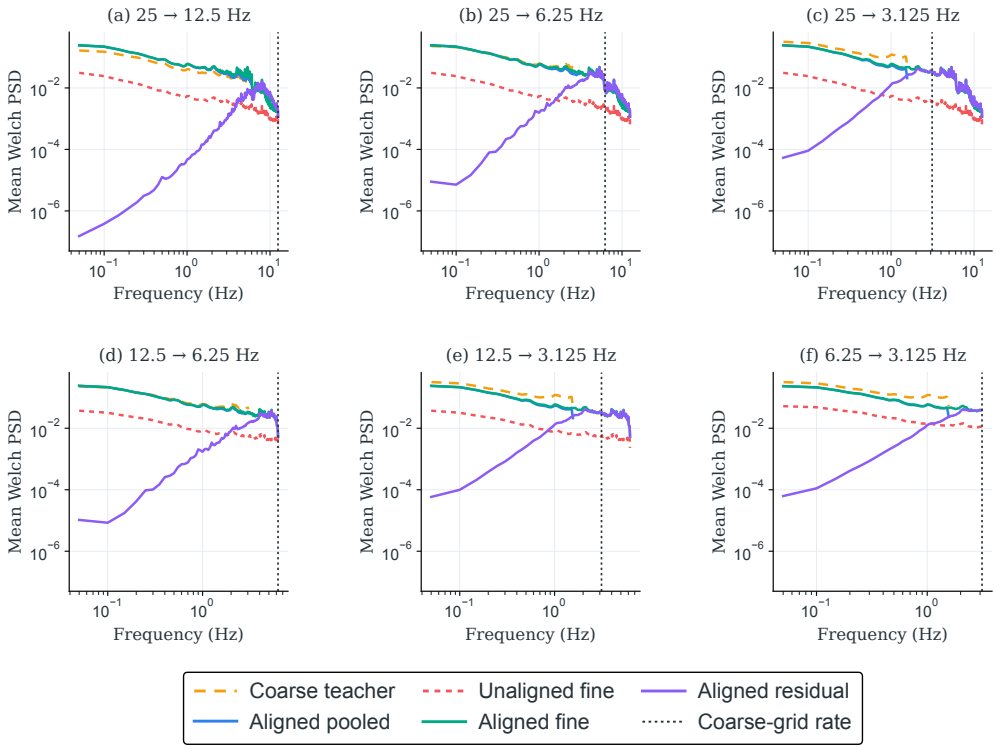


Figure 15: Temporal PSD for the six self-alignment pairs. Frequencies describe variation in the embeddings. Dotted lines mark the coarse-grid rate, not its Nyquist limit; the latter is half that rate.

## G Inference Context and Length Generalization

Our models are trained with physical-time continuous RoPE as a positional information, with the goal of allowing out-of-distribution length generalization. We explore the length generalization of our models trained with fixed-length crops when run over a full track through the lens of structural segmentation performance on full tracks.

We re-encode every Harmonix track with consecutive, non-overlapping windows ranging from 5.12 to 163.84 s, or process the complete track in one pass. The encoder is frozen, and only the surrounding context available to each token changes. We use the student encoder at each of the four native rates and resample its outputs to the same 2-Hz evaluation grid.

For each model, rate, and context, we fit a relational probe on the training split using the model's single ID-selected layer. We fit the joint structure/boundary head at that context on the selected layer as well. The probe therefore sees the same feature-extraction conditions during fitting and evaluation; no reference-context head is transferred to another context. Zero-shot performance omits the learned projection but uses the same ID-selected layer. Both results use the ground-truth number of structural classes and Hungarian label matching.

Finite-window inference pads the terminal window to the requested duration and removes invalid frames. Complete-track one-pass inference instead uses the exact sequence length. We report it separately because it tests length generalization of the encoder and is not mechanically equivalent to a long padded window.

All completed structure conditions include the same 47 test tracks. Shading gives pointwise 95% intervals from 2,000 track-bootstrap resamples, conditional on the selected probe and layer; it does not include training-seed or model-selection uncertainty. The parent and self-aligned sweeps are complete. The control includes all finite contexts and full-track inference at 25 Hz; its other full-track points are unavailable.

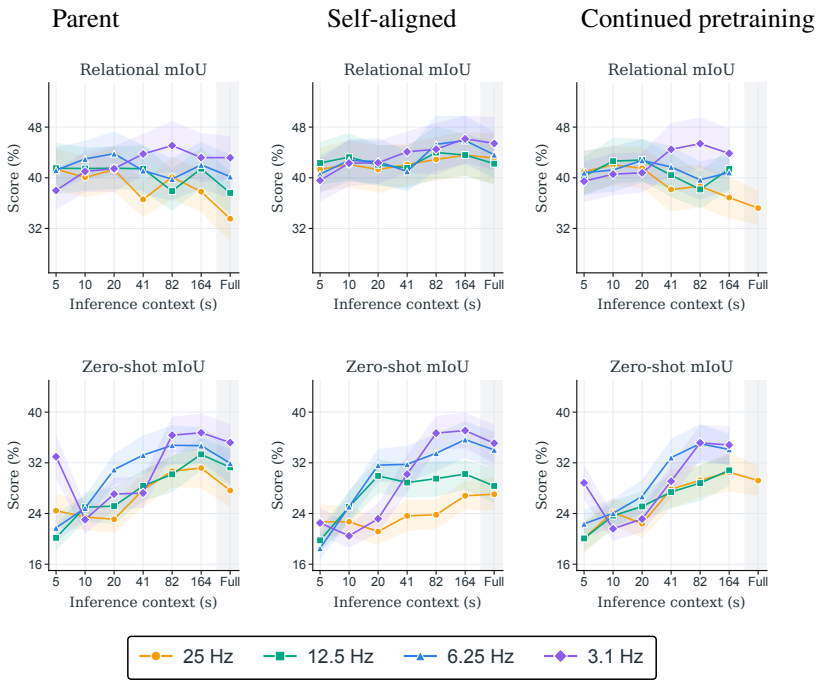


Figure 16: Harmonix relational performance with a probe fitted at each inference context. Zero-shot performance omits the relational projection. Bands: pointwise 95% track-bootstrap intervals. Missing control endpoints remain blank.

For the parent, at 5.12 s, relational mIoU spans 38.0–41.5 across the four queries, whereas zero-shot mIoU spans 20.2–33.0. Fitting the projection removes much of the short-context difference between

rates. The remaining context response is not monotonic: the 6.25-Hz relational query peaks at 20.48 s, while the 3.1-Hz query peaks at 81.92 s. Longer context thus reveals a clearer rate ordering, even after fitting a probe for each condition.

The parent's zero-shot score improves between 5.12 and 81.92 s at every rate, by 3.4–13.0 mIoU points, though individual curves contain reversals. Full-track scores remain above the shortest-context scores but below their best finite-window values. These results support context extrapolation for section clustering with rate-dependent losses, rather than a claim that complete-track inference always gives the best score. The rates also saw different crop lengths during pretraining, so the sweep does not identify which part of that training history causes the difference in extrapolation.

At full-track length, the self-aligned 25-Hz query gains 9.6 relational mIoU points over the parent and 7.9 over continued pretraining. At 163.84 s, where all controls are available, aligned queries exceed their controls by 2.3–6.7 points. These improvements narrow the fine query's long-context disadvantage. They do not imply a uniform improvement in raw geometry: full-track zero-shot mIoU changes by $-0.6$, $-2.9$, $+2.1$, and $-0.1$ points from fine to coarse. The fitted projection remains important for accessing the aligned features' structural information.

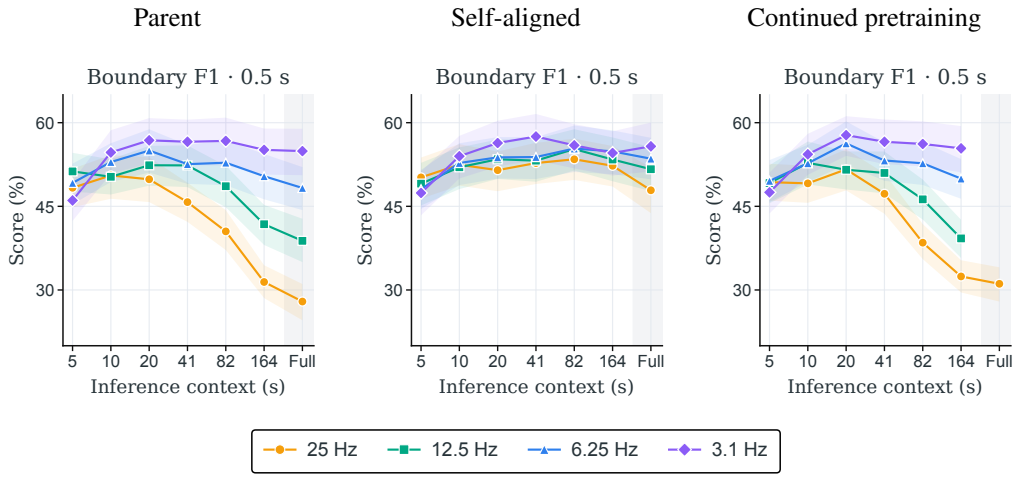


Figure 17: Boundary localization under the matched-context protocol. Self-alignment reduces the fine query's long-context loss. Bands: pointwise 95% track-bootstrap intervals; missing control endpoints remain blank.

Boundary localization is less tolerant of long inputs at the fine end (Figure 17). For the parent, moving from 81.92 s to full-track inference lowers Boundary $F1_{0.5}$ from 40.5 to 27.9 at 25 Hz, but only from 56.7 to 54.9 at 3.1 Hz. After alignment, the 25-Hz query retains 47.9 at full-track length, compared with 31.1 for continued pretraining, while the aligned coarse query reaches 55.8. The aligned fine query still falls from 53.4 at 81.92 s, so longer inference is not cost-free. Because the alignment phase also trains fine queries on longer crops, this comparison measures the combined phase rather than isolating the alignment objective.

# H Intrinsic Geometry Details

## H.1 Resolution-dependent geometry

Similar probe scores can hide differences in how features are organized. We measure TwoNN intrinsic dimension Facco et al. (2017) and RankMe Garrido et al. (2023) to describe variation across track embeddings, and Local vs. Distant Similarity (LDS) Singh et al. (2026a) to describe organization within token grids. We include text-conditioned velocity irreducible variance (VIV), motivated by its strong association with generation quality in visual latent spaces Zhong et al. (2026).

These metrics retain distinct rate-dependent geometry (Figure 18). RankMe is largest at intermediate rates, LDS is strongest at the coarsest rate, and TwoNN varies more mildly. Fusion appears as a horizontal reference because it averages four representations after temporal interpolation instead of defining a fifth token rate. Its TwoNN and RankMe remain within the single-query range.

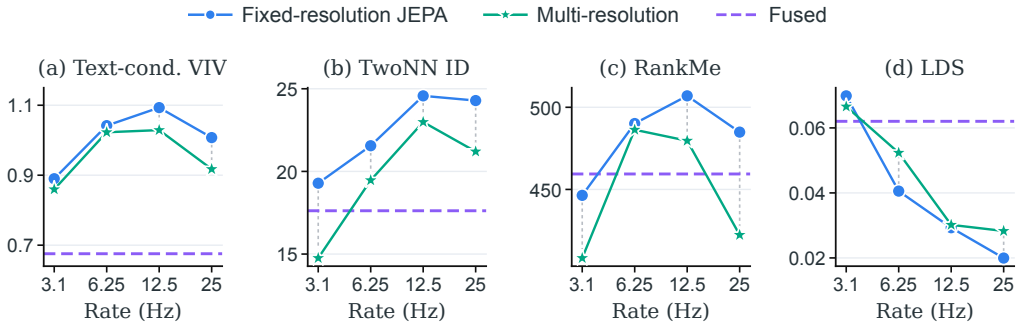


Figure 18: Final-layer geometry across temporal rates. VIV uses caption-conditioned sequences; TwoNN and RankMe use track means, and LDS measures local–distant token similarity. Fusion has no single rate.

## H.2 Velocity irreducible variance

VIV estimates how ambiguous the flow-matching target velocity remains given a noisy latent and its timestep. Lower values indicate less ambiguity under the fitted conditional distribution.

We adapt the Gaussian estimator of Zhong et al. (2026), using the per-dimension convention of their released code. For independent noise $x_0 \sim \mathcal{N}(0, I)$ and a Gaussian target $x_1$ conditioned on group $g$, the path $x_t = (1-t)x_0 + tx_1$ has velocity $x_1 - x_0$. Given target-covariance eigenvalues $\{\lambda_{g,j}\}_{j=1}^{d}$, the conditional velocity ambiguity is

$$\mathrm{VIV}_g(q) = \frac{1}{d}\,\mathbb{E}_{t\sim q}\left[\sum_{j=1}^{d} \frac{\lambda_{g,j}}{(1-t)^2 + t^2\lambda_{g,j}}\right]. \tag{9}$$

For uniform timesteps, this reduces to $\frac{\pi}{2d}\sum_j \sqrt{\lambda_{g,j}}$. We instead use logistic-normal weighting with 128 deterministic quantiles, $t_m = \mathrm{sigmoid}[\Phi^{-1}(0.001 + 0.998(m + 1/2)/128)]$, for $m = 0, \ldots, 127$, where $\Phi$ is the standard-normal CDF. Group scores receive equal weight.

**Sequences and caption conditioning.** We take one 40.96-s crop per track and average frequency positions to obtain $Z_i \in \mathbb{R}^{C\times T}$. Here $C = 1024$ and $T \in \{1024, 512, 256, 128\}$ from finest to coarsest rate; fusion uses the finest grid. Each channel is standardized using its cohort mean and population standard deviation, shared across groups. The covariance describes complete sequences ($d = CT$), not track means. Each group contains a caption anchor and its 32 nearest neighbors. Neighbors are selected by cosine similarity of mean-pooled T5Gemma-2B caption embeddings after subtracting the cohort mean. Figure 18 averages 730 valid caption neighborhoods for each model and rate. These neighborhoods approximate semantic conditioning rather than exact conditioning on a prompt.

**Covariance estimation and uncertainty.** We center each channel–time coordinate within its group and fit $\Sigma_g \approx \Sigma_{T,g} \otimes \Sigma_{C,g}$ with one Kronecker flip-flop iteration, beginning with an identity temporal factor. All pairwise products of factor eigenvalues form the sequence spectrum. We rescale this spectrum to the total sample variance, using denominator $n_g - 1$ for $n_g$ sequences. Whitening floors factor eigenvalues at $10^{-6} \max(\lambda_{\max}, 10^{-12})$. Whiskers show the mean $\pm 1.96$ standard errors across caption neighborhoods; these groups overlap and are not independent replicates. VIV is a diagnostic under Gaussian and separable-covariance assumptions, not a measurement of trained-generator error.

### H.3 Track-level and token-grid geometry

TwoNN and RankMe use full-track mean embeddings. For Figure 18, we process each track in consecutive 4096-frame (40.96-s) windows and discard output positions corresponding to padding. Averaging all valid time–frequency tokens gives one $C$-dimensional vector per track, forming $X \in \mathbb{R}^{n \times C}$ over the 2047-track cohort. We apply no additional channel standardization or vector normalization. Fused and block-pooled views apply their grid operation before this track mean.

**TwoNN intrinsic dimension.** TwoNN estimates the local dimensionality of the track-embedding distribution. We use a maximum-likelihood estimate under the nearest-neighbor ratio model of Facco et al. (2017). For Euclidean distances $r_{i,1}$ and $r_{i,2}$ to the first two non-self neighbors, the estimate is

$$\widehat{d}_{\mathrm{TwoNN}} = \frac{n_v}{\sum_{i \in \mathcal{V}} \log(r_{i,2}/r_{i,1})}, \tag{10}$$

where $\mathcal{V}$ retains the $n_v$ ratios with $r_{i,1} > 10^{-10}$ and $r_{i,2} > r_{i,1}$. We use approximate HNSW search with $M = 32$, construction and query search widths of 200, seed 0, and single-threaded insertion and queries. The estimate describes dimension across track means, not within a token sequence.

**RankMe.** We center each column of $X$ and compute its singular values $s_j$ in float32. Following Garrido et al. (2023), we report

$$\mathrm{RankMe}(X) = \exp\left( -\sum_j p_j \log p_j \right), \qquad p_j = \frac{s_j}{\sum_\ell s_\ell} + 10^{-7}. \tag{11}$$

This measures effective feature rank through the singular-value spectrum, rather than the covariance-eigenvalue spectrum. Larger values indicate that variation is spread across more feature directions.

**Local vs. Distant Similarity.** We adapt the spatial contrast of Singh et al. (2026a) to time–frequency grids. In each window, we uniformly sample up to 256 token positions without replacement using a deterministic seed. For distinct sampled positions $a, b$, let $c_{ab}$ be token cosine similarity and $\delta_{ab} = |t_a - t_b| + |f_a - f_b|$ their Manhattan distance in grid coordinates. With $\tau = \frac{1}{2} \max_{a \neq b} \delta_{ab}$, we compute

$$\mathrm{LDS} = \mathrm{mean}_{0 < \delta_{ab} < \tau} c_{ab} - \mathrm{mean}_{\delta_{ab} \geq \tau} c_{ab}. \tag{12}$$

We return zero if either pair set is empty. Window scores receive equal weight, so longer tracks contribute more windows. For temporal-only sequences, the frequency-distance term is zero. Higher LDS means nearby tokens are more similar than distant ones; distances are measured in token-grid units, not seconds. These geometry metrics have no common better direction and do not measure downstream performance directly.

**Representation autoencoders.** The diffusability diagnostics above concern how easily a latent distribution might be modeled. Using these features for music generation would also require a decoder that preserves useful acoustic detail.

Representation autoencoders pair a frozen semantic encoder with a learned decoder Zheng et al. (2026). Linear separability does not guarantee that a representation retains the low-level information needed for reconstruction Liu et al. (2026), and aggregating earlier layers can trade abstraction for detail Singh et al. (2026b). Music codecs such as SAME optimize their bottlenecks for reconstruction and generation Parker et al. (2026).