# Testing the Continuous Perceptual Trajectory Hypothesis for Musical Motion: A Preregistered Protocol and Open Framework

Feng Yuhan
fengru@gmail.com

September 13, 2026

**Abstract**

We ask whether musical intent ("gesture") can be represented as a continuous latent trajectory over time, and we argue that the central obstacle is not modeling capacity but *identification*. A representation trained on top of assumed gesture-preserving transformations can trivially "succeed" and still carry no musical-motion information, because the transformations, the notion of a negative gesture, the alignment convention, and the evaluation metric are all researcher-defined. We formalize six threats to validity (T1–T6) and derive, for each, a concrete mitigation. We then present a framework whose design is forced by those mitigations: transformations are treated as a *parameterized hypothesis* whose semantics are established by human ratings before use as positives; variable-length trajectory distances are reported in raw, length-normalized, and dynamic-time-warped forms, with the normalized form primary; representation collapse is an independent pass/fail criterion, enforced during training by VICReg-style variance and covariance terms; continuity is expressed as multi-step predictability, never as an $L_2$ penalty on adjacent latents; and success is a conjunction of criteria that explicitly includes losing to stronger baselines. We release an open-source implementation and a pilot on 40 solo-piano excerpts from MAESTRO (160 segments, 1,600 controlled variants) plus a 400-trial human rating study that fixes the transformation ontology at the `(transform, parameter)` level, allowing a perceptual boundary to be estimated rather than a binary label asserted. We report only engineering validation here.

## 1 Introduction

### 1.1 Motivation

Systems that generate, accompany, or edit music increasingly need a handle on high-level musical intent: not "which notes" or "which waveform," but *how the music moves*—whether a phrase pushes upward, relaxes, accelerates, or resolves. A natural internal object for such intent is a *continuous perceptual trajectory*: a time-indexed latent state

$$z_{1:T} = f(x_{1:T}), \qquad z_t \in \mathbb{R}^d, \tag{1}$$

where $x_{1:T}$ is an audio excerpt and a *gesture* is a temporally localized segment of the trajectory,

$$G = \{z_t\}_{t=a}^{b}, \qquad 0 \le a < b \le T. \tag{2}$$

This representation promises continuity (nearby states are related), compositionality (gestures concatenate), editability (modify a sub-trajectory), and prediction (extrapolate the trajectory). The

idea resonates with a long tradition treating musical gesture as sound-related movement (Godøy and Leman, 2010; Meyer, 1956).

### 1.2 The trajectory hypothesis

We make the hypothesis explicit and falsifiable. Let $f$ be an encoder, $T$ a transformation family, and $d$ a distance on trajectories. Define the *realization set* of an excerpt $x$ under a set of realizational transformations $\mathcal{T}_{\text{preserve}}$ as

$$\mathcal{R}(x) = \{\, T(x) \;:\; T \in \mathcal{T}_{\text{preserve}} \,\}. \tag{3}$$

The trajectory hypothesis is the conjunction of four sub-hypotheses:

- **H1 (perceptual meaningfulness).** There exists an encoder $f$ such that $d$ in $z$-space predicts human judgments of *perceived musical motion similarity* better than strong acoustic and self-supervised baselines.
- **H2 (invariance to realization).** For $T$ that humans agree preserves motion, $d(f(x), f(T(x)))$ is small relative to $d(f(x), f(y))$ for a genuinely different $y$.
- **H3 (predictability).** $z_{1:T}$ supports multi-step prediction better than extrapolation or frozen-feature baselines, i.e., the trajectory carries structure beyond local smoothness.
- **H4 (local editability).** A localized error $e_t = d(z_t, \hat{z}_t)$ identifies an interval whose regeneration corrects an intent error without collateral change elsewhere.

H4 is deliberately last: it is meaningful only if H1–H3 hold. The framework therefore defers generation and editing and treats them as downstream consequences rather than as evidence.

### 1.3 Contributions

1. An *identification analysis* (Section 3): six threats that inflate gesture claims, with worked shortcut constructions showing that each mitigation alone is insufficient (Section 3.1).
2. A *parameterized human ontology* (Section 4.3): a rating protocol and estimator that classify each (transform, parameter) setting into `preserving`/`altering`/`ambiguous`/`uncertain` with agreement gates, estimating a perceptual boundary (e.g., in tempo) instead of asserting a binary status.
3. A *preregistered criterion set and statistical plan* (Section 5): success is a conjunction (C1–C5) that includes losing to MERT/CLAP/handcrafted as an explicit failure condition.
4. An *open framework and pilot* (Sections 6–9).

### 1.4 Related work

**Music and audio representation learning.** MERT (Li et al., 2024) and MusicFM (Won et al., 2024) learn general-purpose music embeddings by masked acoustic modeling; CLAP (Elizalde et al., 2023; Wu et al., 2023b) aligns audio with text; CLaMP (Wu et al., 2023a) aligns symbolic music with text. These provide the strong frozen features our human-alignment criterion must beat.

**Invariance-based audio SSL.** COLA (Saeed et al., 2021), BYOL-A (Niizumi et al., 2021), and VICReg (Bardes et al., 2022) learn representations stable under augmentation. They are close to H2 but are typically evaluated on downstream classification, not on a human-defined motion geometry, and they do not treat transformation semantics as an empirical object.

**Predictive representation learning.** Joint-embedding predictive architectures (I-JEPA (Assran et al., 2023); V-JEPA (Bardes et al., 2024)) combine invariance with prediction in latent space. Our framework is compatible with these objectives but is agnostic to the encoder; the contribution is the protocol that decides whether *any* representation can be said to capture musical motion.

**Gesture and perception.** The gesture literature (Godøy and Leman, 2010; Meyer, 1956) motivates the construct; our contribution is to operationalize it so that it can be measured and, importantly, falsified.

### 1.5 Scope and non-goals

We restrict the pilot to **solo piano** to control timbre, polyphony, and vocal semantics so that transformation semantics can be established cleanly. We explicitly do *not* propose a generative model, decoder, or editing system in this phase; those are downstream and admissible only if H1–H3 hold. We also do not claim that "gesture" is the right ontology—the framework is designed so that this can come out false.

## 2 Setup and notation

### 2.1 Audio and segmentation

Let $x \in \mathbb{R}^N$ be a mono waveform at rate $s$. A *segmenter* produces candidate excerpts $S = \{x^{(i)}\}$ by fixed-window slicing at durations $\mathcal{D} = \{d_1, \dots, d_k\}$ with hop $h = \rho d$ (we use $\rho = 1/2$). A window is kept if its RMS level in dBFS exceeds $\ell_{\min}$:

$$\mathrm{dB}(x^{(i)}) = 20 \log_{10}\left( \sqrt{\tfrac{1}{|x^{(i)}|} \textstyle\sum_n (x_n^{(i)})^2} + \epsilon \right) > \ell_{\min}. \tag{4}$$

Fixed-window segmentation is deliberate: it avoids importing an unvalidated gesture-segmentation model into the ontology (T1). Human segmentation is deferred.

### 2.2 Gesture as a trajectory segment

A frame encoder $f_\phi$ maps audio to a sequence; a temporal encoder $g_\theta$ maps the sequence to a low-dimensional trajectory:

$$z_{1:T} = g_\theta(f_\phi(x)), \qquad f_\phi \text{ frozen}, \quad z_t \in \mathbb{R}^d,\ d \ll \dim f_\phi. \tag{5}$$

The bottleneck $d$ is part of H3: a representation that memorizes frame features can predict trivially without learning structure.

### 2.3 Transformations and realization

A transformation $T_\psi$ maps audio to audio. A *setting* is a pair $(\text{transform}, \psi)$; write $\mathcal{T}$ for the set of settings. The partition $\mathcal{T} = \mathcal{T}_{\text{preserve}} \sqcup \mathcal{T}_{\text{alter}} \sqcup \mathcal{T}_{\text{ambiguous}}$ is *not assumed*; it is estimated in Section 4.3.

### 2.4 Distance on variable-length trajectories

Tempo-like transforms change $T$, so $d$ must specify an alignment. For $a = a_{1:T_a}$, $b = b_{1:T_b}$:

$$d_{\text{raw}}(a,b) = \frac{1}{\min(T_a,T_b)} \sum_{t=1}^{\min(T_a,T_b)} \|a_t - b_t\|_2, \tag{6}$$

$$d_{\text{norm}}(a,b) = \frac{1}{L}\sum_{t=1}^{L}\|\tilde{a}_t - \tilde{b}_t\|_2, \quad \tilde{a} = R(a,L),\ \tilde{b} = R(b,L), \tag{7}$$

$$d_{\text{dtw}}(a,b) = \frac{1}{|W|}\min_{W} \sum_{(i,j)\in W} \|a_i - b_j\|_2, \tag{8}$$

where $R(\cdot, L)$ linearly resamples to length $L$ on $[0,1]$. The normalized distance $d_{\text{norm}}$ is primary: it tests invariance to *rate* without letting dynamic time warping absorb genuine differences (T2). $d_{\text{dtw}}$ (Sakoe and Chiba, 1978) is a robustness check. Distances use per-dimension standardized features when `normalized` is on, so no single high-variance dimension dominates.

### 2.5 Model geometry and RSA

The representational dissimilarity matrix (RDM) of embeddings $\{e_i\}$ is the condensed vector $\text{RDM}_{\text{model}} = \text{pdist}(\{e_i\}, d)$. Model and human geometry are compared by Spearman correlation of condensed RDMs (Kriegeskorte et al., 2008),

$$\text{RSA} = \rho_{\text{Spearman}}\big(\text{RDM}_{\text{model}}, \text{RDM}_{\text{human}}\big), \tag{9}$$

and by triplet accuracy $\mathbb{P}(d(a,p) < d(a,n))$, which assumes no parametric form.

## 3 The identification problem

Every component of the test—positives, negatives, alignment, and metric—is chosen by the researcher. We give six threats and a construction showing that mitigations are not independent.

**T1: transformation semantics are assumed, not measured.** Suppose we declare $\mathcal{T}_{\text{preserve}} = \{\text{transposition}, \text{gain}, \text{brightness}\}$. Consider the encoder $f_{\text{pc}}(x) = \text{pitch-class histogram}(x)$, which discards onset timing, contour shape, and duration. It is *exactly* invariant to transposition, gain, and timbre, so $d(f_{\text{pc}}(x), f_{\text{pc}}(T(x))) = 0$ for all $T \in \mathcal{T}_{\text{preserve}}$—while being unable to represent motion at all. The failure is the test, not the encoder.

**T2: variable-length trajectories.** With $d_{\text{norm}}$, a temporally warped trajectory can be made arbitrarily close by interpolation; with a flexible $d_{\text{dtw}}$, a slowly rising contour can align to a fast one. The same representation can pass or fail H2 depending on the metric.

**T3: collapse.** Let $f_c(x) = c$ for a constant $c$. Then $d(f_c(x), f_c(y)) = 0$ and $d(f_c(x), f_c(T(x))) = 0$ for all $x, y, T$. $f_c$ satisfies H2 perfectly and minimizes any continuity penalty, yet its RDM is all zeros and its effective rank is zero. An invariance number reported alone is meaningless.

**T4: trivial prediction.** If $z_t \approx z_{t-1}$, persistence predicts the next step almost perfectly, so one-step prediction is uninformative. Prediction must be multi-step and compared against strong extrapolation baselines.

| Threat | Mitigation in this work |
|---|---|
| T1 | Human rating classifies each (transform, parameter); unvalidated settings are excluded from decisions |
| T2 | Distances reported as raw, normalized (primary), and DTW (robustness only) |
| T3 | Mandatory VICReg variance/covariance; collapse is an independent pass/fail criterion (C1) |
| T4 | Multi-step prediction with persistence, linear, ridge-on-frozen, constant-mean baselines (C5) |
| T5 | Architecture and objective ablations over a fixed frozen encoder; baselines reported as-is |
| T6 | A single named construct with a fixed prompt and agreement statistics |

Table 1: Threats to validity and their mitigations.

**T5: unfair baselines.** A learned encoder trained with invariance and prediction is compared to a frozen model trained for another objective; any gain may come from the objective, not the representation.

**T6: construct ambiguity.** "Are these two similar?" conflates motion with emotion, tension, loudness, tempo, or timbre. Low agreement then reflects an under-specified construct, not perceptual subtlety.

### 3.1 Why a conjunction

The constructions are not independent. A constant encoder defeats invariance and continuity; a pitch-class histogram defeats invariance while losing motion; a smooth representation passes one-step prediction; a flexible metric rescues a bad representation. No single mitigation removes all shortcuts, so the claim is only meaningful as a conjunction (C1–C5), and the framework surfaces any single failure explicitly.

## 4 Framework

### 4.1 Segmentation

Durations $\mathcal{D} = \{1, 2, 4\}$ s ($\{0.5, 1, 2, 4, 8\}$ supported), hop ratio $\rho = 1/2$, target rate 22.05 kHz, $\ell_{\min} = -50$ dBFS, and at most $m$ segments per source (pilot $m = 4$). Candidates are filtered *before* any file is written, so the pilot produces $\mathcal{O}(\#\text{sources} \times m)$ segments rather than $\mathcal{O}(\#\text{windows})$.

### 4.2 Parameterized transformations

We apply four families with explicit grids (Table 2). Each setting is independent; the hypothesis column is a prior only. `brightness` is a high-shelf filter ($y = y_{\text{low}} + g(y - y_{\text{low}})$ with a one-pole lowpass). Articulation changes are intentionally absent because they require note-level (MIDI) edits; this is recorded as a known gap, not approximated.

### 4.3 Human ontology validation

**Task.** For each anchor $A$ and variant $B = T(A)$, a rater answers one question on a 7-point scale: "A and B are two versions of the same musical excerpt. Rate how similar $B$'s musical motion is to $A$'s, from 1 (very different) to 7 (same motion)." The prompt is fixed and names exactly one construct (T6). Anchors and variants are presented in a static HTML page with in-page audio, so the study needs no server and works from `file://`.

| Transform | Family | Parameter grid | Prior |
|---|---|---|---|
| `pitch_shift` | pitch | semitones $\{+2, -3\}$ | preserving |
| `gain` | dynamics | dB $\{+6, -6\}$ | preserving |
| `brightness` | timbre | high-shelf dB $\{+9, -9\}$ | preserving |
| `time_stretch` | tempo | rate $\{0.8, 0.9, 1.1, 1.25\}$ | ambiguous |

Table 2: Parameterized transforms. A transform with disagreeing settings is labeled `mixed`.

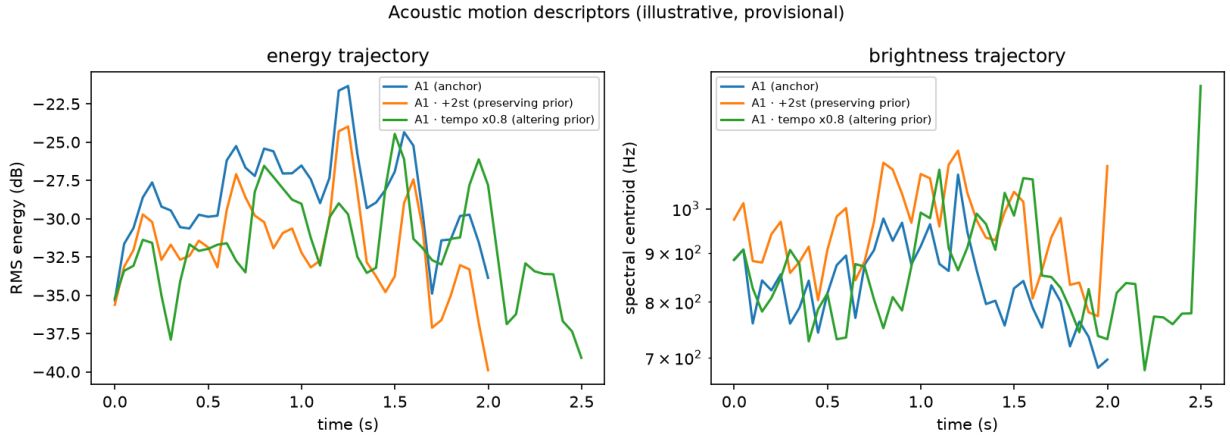


Figure 1: Acoustic motion descriptors for one anchor and two of its realizations (illustrative; provisional ontology). Left: short-time RMS energy over time; right: spectral centroid (brightness) over time. Transposition (`+2st`) preserves the energy shape and raises the brightness curve; the strong tempo change (`x0.8`) preserves the shape while changing the duration. This motivates the trajectory view but is not a representation result.

**Aggregation.** Responses are grouped by setting $k = (\text{transform}, \psi)$. For each $k$ we compute the mean, a bootstrap 95% CI (Efron and Tibshirani, 1993), the per-response category distribution

$$\pi_k(c) = \frac{1}{|R_k|} \sum_{r \in R_k} \mathbf{1}[\text{cat}(r) = c], \qquad c \in \{\text{preserving}, \text{ambiguous}, \text{altering}\}, \tag{10}$$

where $\text{cat}(r) = \text{preserving}$ if $r \geq \text{mid} + \text{margin}$, altering if $r \leq \text{mid} - \text{margin}$, else ambiguous, and agreement via Krippendorff's ordinal $\alpha$ (Krippendorff, 2004) and, for complete designs, Fleiss' $\kappa$ (Fleiss, 1971).

**Classification.** With $\text{mid} = 4.0$, $\text{margin} = 0.5$, minimum raters $R_{\min}$, minimum trials $N_{\min}$, and minimum agreement $\alpha_{\min}$:

$$\text{status}(k) = \begin{cases} \text{uncertain} & \text{if } n_{\text{raters}} < R_{\min} \text{ or } n_{\text{trials}} < N_{\min} \text{ or } \alpha_k < \alpha_{\min}, \\ \text{preserving} & \text{else if } \text{CI}_k^{\text{lo}} \geq \text{mid} + \text{margin}, \\ \text{altering} & \text{else if } \text{CI}_k^{\text{hi}} \leq \text{mid} - \text{margin}, \\ \text{ambiguous} & \text{otherwise.} \end{cases} \tag{11}$$

A transform is `mixed` if its settings disagree. The ontology records raw distributions, so downstream choices are auditable. Two properties matter: `uncertain` settings are excluded from the

invariance decision, and the ontology is computed *before* any representation is evaluated, breaking the circularity of using model output to define the construct.

### 4.4 Representations

We compare four frozen encoders and one learned model: (1) *handcrafted*: per-frame RMS, spectral centroid, spectral bandwidth, zero-crossing rate, onset strength, and $f_0$ (`pyin`), z-scored per dimension; (2) *log-mel*: 64-bin frames at a fixed hop; (3) *frozen MERT* (Li et al., 2024): mid-layer hidden states mean-pooled to a fixed rate; (4) *frozen CLAP* (Elizalde et al., 2023; Wu et al., 2023b): windowed audio embeddings; and (5) a *learned trajectory*: a frozen frame encoder, then a 1D temporal transformer (Vaswani et al., 2017) with sinusoidal positional encoding, $d_{\text{model}} = 128$, two layers, four heads, and a projection to $z_t \in \mathbb{R}^{64}$. The small learned model is deliberate: H1–H3 concern the geometry and dynamics of the representation, not encoder capacity, and a small model makes architecture/objective ablations attributable (T5).

### 4.5 Distances and geometry

As defined in Sections 2.4–2.5. For invariance (E4) we compare, per setting, $d(z_A, z_{T(A)})$ against cross-excerpt distances $d(z_A, z_B)$, and summarize by the AUC of the preserving-vs-other split and the mean gap. For human alignment (E3) we report RSA and triplet accuracy.

### 4.6 Objectives

Let $z = g_\theta(f_\phi(x))$ and $z^+ = g_\theta(f_\phi(T(x)))$ for a setting validated as preserving. The loss is

$$\mathcal{L} = \lambda_{\text{var}}\ell_{\text{var}} + \lambda_{\text{cov}}\ell_{\text{cov}} + \lambda_{\text{inv}}\ell_{\text{inv}} + \lambda_{\text{pred}}\ell_{\text{pred}} + \lambda_{\text{tri}}\ell_{\text{tri}}. \tag{12}$$

With $\gamma = 1$ and $C = \text{Cov}(z)$,

$$\ell_{\text{var}}(z) = \frac{1}{d}\sum_{j=1}^{d} \max\left(0, \gamma - \sqrt{\text{Var}(z_{\cdot j}) + \epsilon}\right), \tag{13}$$

$$\ell_{\text{cov}}(z) = \frac{1}{d}\sum_{i \neq j} C_{ij}^2, \tag{14}$$

$$\ell_{\text{inv}}(z, z^+) = \|z - z^+\|_2^2, \tag{15}$$

$$\ell_{\text{pred}} = \frac{1}{\sum_i w_i}\sum_{i=1}^{k} w_i \, \|z_{t+i} - \hat{z}_{t+i}\|_2^2, \quad w_i = 1/i, \tag{16}$$

$$\ell_{\text{tri}} = \max(0, \|z_a - z_p\|_2 - \|z_a - z_n\|_2 + m)\,. \tag{17}$$

Continuity is implemented *only* through $\ell_{\text{pred}}$. We do not add a term penalizing $\|z_t - z_{t+1}\|$, because that term is minimized by the constant encoder of T3; predictability is the correct expression of continuity. The variance/covariance terms follow VICReg (Bardes et al., 2022) and the alignment–uniformity view (Wang and Isola, 2020); the triplet term follows metric learning (Schroff et al., 2015).

| ID | Experiment | Question | Criterion |
|---|---|---|---|
| E1 | Transformation validation | Which settings preserve motion? | input to all |
| E2 | Collapse diagnostics | Is the representation non-degenerate? | C1 |
| E3 | Human similarity (RSA/triplet) | Does model geometry match humans? | C4 |
| E4 | Invariance | Are preserving settings close? | C2 |
| E5 | Specificity | Are altering settings separated? | C3 |
| E6 | Multi-step prediction | Structure beyond smoothness? | C5 |
| E7 | Architecture ablation | transformer vs. GRU vs. mean-pool | attribution |
| E8 | Objective ablation | contribution of var/cov/inv/pred/tri | attribution |

Table 3: Experiment matrix.

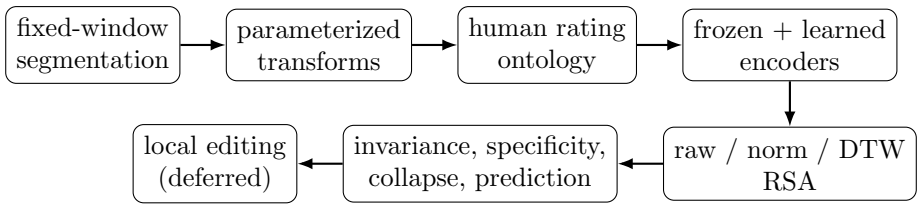


Figure 2: Phase 1 pipeline. Transformation semantics are fixed by humans before any representation is evaluated; editing is downstream of the preregistered criteria.

### 4.7 Experiments

### 4.8 Local editing (deferred)

H4 requires an inverse map from trajectory to audio—the generative component we have not built—so editability cannot yet be tested. A decoder-free proxy is possible (retrieval: does the nearest actual segment realize an edited $z$-segment?), but it is not equivalent to editing and is not counted as evidence for H4.

## 5 Preregistered criteria and statistical plan

A representation is reported as evidence only if *all* criteria hold on held-out data.

- **C1 (health).** $\text{mean}_j \operatorname{Var}(z_{\cdot j}) > v_{\min}$ and $\operatorname{erank}(z) > r_{\text{frac}}\, d$, with $v_{\min} = 10^{-3}$, $r_{\text{frac}} = 0.1$, and $\operatorname{erank}(z) = \exp H(\sigma/\|\sigma\|_1)$ over singular values $\sigma$.
- **C2 (invariance).** For every setting with status `preserving`, $\text{mean}\, d(z_A, z_{T(A)}) < \text{mean}\, d(z_A, z_B)$ over different excerpts, and separation AUC $> 0.5$ under $d_{\text{norm}}$.
- **C3 (specificity).** Settings with status `altering` yield distances clearly above preserving ones (same separation analysis).
- **C4 (human alignment).** Triplet accuracy exceeds MERT, CLAP, and handcrafted features under the same distance.
- **C5 (prediction).** Multi-step MSE is lower than persistence, linear extrapolation, ridge on frozen features, and constant mean.

**Decision rule.** If C1 fails, no other claim is made. If C2 fails for a `preserving` setting, it is reclassified `ambiguous` and the protocol re-run (a data-driven ontology update, not a metric

change). If C4 fails, the gesture ontology is reported as *unsupported at this stage*, and the failure documented rather than hidden.

**Statistical plan.** Agreement is reported per setting. Invariance uses bootstrap CIs over excerpt pairs; human alignment uses bootstrap CIs over triplets. For E7–E8 we report paired bootstrap intervals across seeds. Because many settings are tested, we report the full family rather than a single best setting, and we never select on the test split.

# 6 Pilot

**Data.** Forty solo-piano excerpts from MAESTRO (Hawthorne et al., 2019) (train split), composer-stratified, totaling 334 minutes. We selected a mirror storing extracted `.wav` files that match MAESTRO's official metadata paths; audio is not redistributed and only the subset manifest is versioned.

**Stimuli.** Segmentation at $\{1, 2, 4\}$ s, 50% hop, at most four segments per excerpt, 22.05 kHz mono: **160 segments**; parameterized transforms yield **1,600 variants** (10 settings per segment).

**Human study.** A dependency-free static HTML page presents **400 rating trials** (40 per setting) with in-page audio and JSON export; multiple raters are merged before scoring. The minimum is five raters for an ontology to count as validated.

**Baselines and models.** Handcrafted acoustic features, log-mel, frozen MERT, frozen CLAP (weights not redistributed), and the learned trajectory. Compute is a single CPU workstation.

**Artifacts.** `segments.jsonl`, `variants.jsonl`, `transformation_validation.html`, a trials JSON, per-rater response JSONs, and `ontology.json`.

# 7 Preliminary engineering validation (not evidence)

This section reports pipeline correctness only; all numbers use the provisional priors of Table 2 and are *not* evidence about musical motion. The unit-test suite passes on synthetic audio (segmentation and transforms; metrics including Fleiss' $\kappa$, Krippendorff's $\alpha$, RSA, and collapse diagnostics; losses; distances; the rating and triplet flows; experiment drivers): 30 tests. On the 160 real segments and 1,600 variants, the pipeline generates the 400-trial study and runs the frozen baselines without error. Under the provisional ontology, on a 40-segment subset, log-mel frames give separation AUC 0.973 and handcrafted features 0.999, with healthy collapse diagnostics for both. These approach the ceiling that *any* reasonable representation reaches under an assumed ontology; they quantify the inflation described by T1 and are the strongest argument for making the human ontology the blocking step. We intentionally do not interpret them.

# 8 Limitations and threats to validity

The central limitation is that the human ontology has not been collected, so H1–H4 are untested and every invariance number is provisional. The timbre transform is a high-shelf proxy, and articulation changes are absent because they need note-level (MIDI) edits. The pilot is solo piano by design; generalization to polyphony, voice, and multi-instrument music is untested and may change what "motion" means. Frozen MERT and CLAP are implemented but were not run in the pilot, so the human-alignment comparison is incomplete. The distance choice and the classification margin are conventions; the ontology schema records raw distributions so they can be probed post hoc. Fixed-window segmentation may bias the candidate gestures; it is chosen to protect the ontology, not because it is optimal.

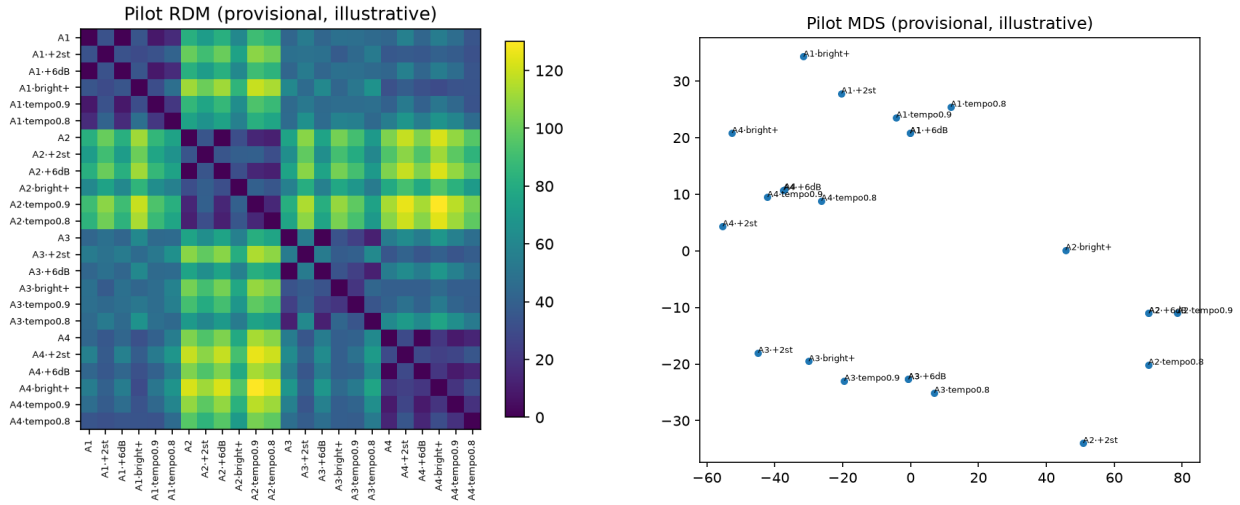


Figure 3: Latent geometry of the pilot under the provisional ontology (log-mel pooled embeddings; illustrative, not evidence). Left: pairwise distance matrix over four anchors (`A1`–`A4`) and their variants. Right: metric MDS of the same embeddings. Within-anchor variants form blocks and clusters, while cross-anchor distances are larger, and the two tempo settings span a wide range—consistent with the `mixed` prior and with the need to estimate the tempo boundary from humans.

## 9 Open-source release and reproducibility

The repository contains segmentation and transforms, the human-study runner, parameterized ontology scoring, metrics, representation and baseline encoders, losses, experiment drivers, visualization, and the preregistered criteria and protocol, and is available at `https://github.com/Fengrru/mmt`. Reproduction is a short sequence of commands (`docs/reproduction.md`). Code is MIT-licensed; MAESTRO audio is subject to its own license (CC BY-NC-SA 4.0) and is not redistributed. The release includes a `CONTRIBUTING.md` whose first rule is that model results must never define the ontology, preserving the identification argument for downstream contributors.

## 10 Conclusion

We argued that the difficulty in learning a musical-motion trajectory is identification, not capacity, and converted that concern into a preregistered protocol and an open implementation. The framework derives its design from six concrete threats and is constructed so that a degenerate shortcut fails at least one mandatory criterion. Its value does not depend on the hypothesis being true: if human ratings show the assumed transformations do not preserve motion, or if a simple acoustic baseline already matches the human construct, the framework reports this cleanly and early, before any generative system is built. We release the pilot and protocol so the human study can be run and, if warranted, the hypothesis falsified.

## A Ontology schema

```
{
  "validated": true,
```

```
  "scale": 7, "midpoint": 4.0, "margin": 0.5,
  "entries": [{
    "key": "time_stretch:rate=0.9",
    "transform": "time_stretch",
    "params": {"rate": 0.9},
    "n_ratings": 48, "n_raters": 6, "n_trials": 8,
    "mean": 6.0, "ci_low": 5.7, "ci_high": 6.3,
    "judgment": {"preserving": 1.0, "ambiguous": 0.0, "altering": 0.0},
    "agreement": {"krippendorff_alpha_ordinal": 0.67, "fleiss_kappa": 0.6},
    "status": "preserving"
  }],
  "by_key": {"time_stretch:rate=0.9": "preserving"},
  "by_transform": {"time_stretch": "mixed"}
}
```


# B Latent shared projection (provisional)

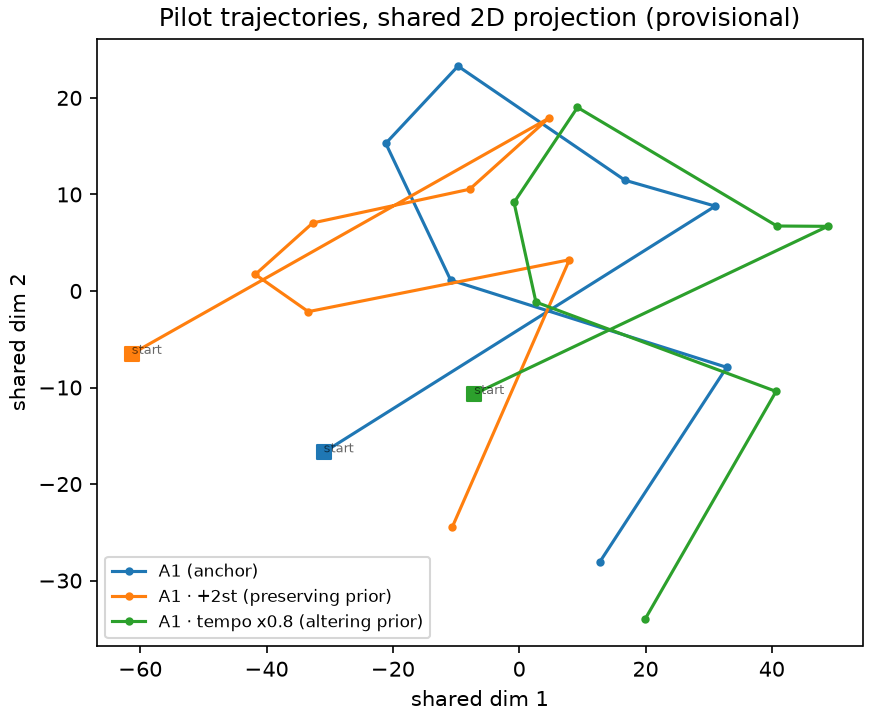


Figure 4: Shared 2D PCA projection of one anchor and two realizations (log-mel frames, coarse to eight time steps; illustrative). A single projection basis is fit on all frames so trajectories are comparable. The high-dimensional frame trajectory is deliberately not used as a primary figure; the geometry of representation is summarized by the RDM/MDS (Figure 3), while interpretable motion is shown in Figure 1.

## C Reporting checklist

For every reported number state: (i) the distance used; (ii) the ontology status of every setting involved; (iii) agreement and $n$; (iv) whether `validated` is true; (v) collapse diagnostics; (vi) all baselines. A number that omits any of these is not part of the claim.

## References


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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

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. *arXiv preprint*, 2024.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.

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

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

Rolf Inge Godøy and Marc Leman, editors. *Musical Gestures: Sound, Movement, and Meaning*. Routledge, 2010.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations (ICLR)*, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage, 2nd edition, 2004.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, et al. MERT: Acoustic music understanding model with large-scale self-supervised training. In *International Conference on Learning Representations (ICLR)*, 2024.

Leonard B. Meyer. *Emotion and Meaning in Music*. University of Chicago Press, 1956.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. BYOL for audio: Self-supervised learning for general-purpose audio representation. In *International Joint Conference on Neural Networks (IJCNN)*, 2021.

Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, 2020.

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

Shangda Wu, Dingyao Yu, Xu Tan, and Maosong Sun. CLaMP: Contrastive language-music pretraining for cross-modal symbolic music information retrieval. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2023a.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023b.