MAJEPPA: MORPHING AND ASSESSING IN A UNIFIED PIANO PERFORMANCE SPACE

Jinwen Zhou[1] **Huan Zhang**[1] **Weixi Zhai**[2]
Jinhua Liang[1] **Aidan O. T. Hogg**[1] **Simon Dixon**[1]
[1] Queen Mary University of London, UK
[2] Quanzhou Normal University, China
[jinwen.zhou, huan.zhang, jinhua.liang, a.hogg, s.e.dixon]@qmul.ac.uk[1], wishzhai@gmail.com[2]

ABSTRACT

We present MAJEPPA, a self-supervised framework to learn piano performance representations that span the full skill spectrum, from beginner practice sessions to virtuoso concert recordings. We curate the MAJEPPA dataset, comprising ∼4,000 annotated recordings across six expertise levels and six recording contexts. We adapt a single pre-trained MIDI autoregressive model with a joint objective: next-token prediction learns score-conditioned performance generation at various skill levels, while InfoNCE and supervised contrastive losses align abstract score and performance representations in a joint embedding space. The proposed model both generates and understands performances in a unified framework. By introducing the EVPMR benchmark, a suite of downstream tasks spanning quality assessment, competition ranking, mistake and technique classification, we evaluate the learnt representations, demonstrating progress towards a real-world model for the piano performance space.

1. INTRODUCTION

Piano performance is a remarkably rich domain of human expression. The same score can be realised as a child's halting first attempt, an adult student sight-reading, a teacher demonstrating phrasing at half tempo, or a virtuoso commanding the concert stage. Understanding this variation computationally, distinguishing not just *what* is played but *how*, is central to music education, where intelligent systems must diagnose mistakes [1–3], assess quality [4, 5], provide formative feedback [6], estimate difficulty [7, 8], and personalise learning [9] across student abilities. Recent efforts have begun capturing the learning process itself through rehearsal datasets [10] and multimodal practice recordings [11, 12]. Yet existing approaches address these goals in isolation: task-specific models are trained on narrow datasets for individual objectives, yielding representations that do not transfer across tasks or skill levels.

© J. Zhou, H. Zhang, W. Zhai, J. Liang, A. O. T. Hogg, and S. Dixon. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** J. Zhou, H. Zhang, W. Zhai, J. Liang, A. O. T. Hogg, and S. Dixon, "MAJEPPA: Morphing and Assessing in a Unified Piano Performance Space", in *Proc. of the 27th Int. Society for Music Information Retrieval Conf.,* Abu Dhabi, UAE, 2026.

We argue that a *unified* representation of piano performance is more foundational, encoding core performance parameters from which higher-level structure is derived, and more useful in practice. However, building such a representation is hindered by two gaps in MIR research. First, a **data gap**: existing symbolic music datasets focus almost exclusively on professional performances [13, 14], leaving most real-world piano playing (practice, sight-reading, demonstrations) unrepresented. Second, a **methodological gap**: we need a learning framework that captures *abstract* performance characteristics (style, accuracy, expressivity) alongside and linked to the score. We address this gap with the Joint-Embedding Predictive Architecture (JEPA) [15], a framework that learns representations by predicting in latent space, focusing on high-level semantics. Furthermore, LLM-JEPA [16] demonstrates that predictive latent modelling with language models can yield representations that are not only abstract but also perceptually grounded.

Given a score, we want to predict the *abstract character* of how that score will be realised. The score serves as the context view (specifying *what* to play) and the performance as the target view (capturing *how* it is played). The model must predict the abstract performance embedding from the score, along with descriptive conditioning (performer type, recording context), without observing the performance tokens. A joint generative objective prevents representation collapse and retains the model's ability to generate, resulting in a single model that both produces performances and yields transferable embeddings that specialise in understanding real-world performances.

To close the data gap, we curate the **MAJEPPA dataset**, comprising ∼4,000 annotated piano recordings spanning six expertise levels and six recording contexts. We evaluate our JEPA-based model on the Evaluation of Piano MIDI Representations (EVPMR) benchmark using frozen representations with linear probes. Our contributions are:

1. The **MAJEPPA dataset**, the first large-scale dataset to span the full performance spectrum from child beginners to virtuosi;
2. The first **JEPA** to model for joint symbolic piano-performance generation and representation learning, using score–performance pairs as a natural two-view structure;
3. The **EVPMR** benchmark suite, designed for evalua-

tion across downstream tasks, including Chopin competition ranking, technique detection, performance quality regression, and conspicuous mistake prediction, where our models outperform various baselines.

2. RELATED WORK

2.1 Music and MIDI representation learning

Approaches to learning representations from symbolic music are shaped by how MIDI is serialised into a sequence, and tokenisation schemes such as REMI [17], Compound Word [18], OctupleMIDI [19], and PerTok [20] provide the input vocabularies on top of which the following families of self-supervised objectives are applied.

Masked-reconstruction pre-training: MusicBERT [19] and MidiBERT-Piano [21] apply BERT-style masked token prediction to symbolic music, producing representations that transfer to composition-level classification (genre, emotion, melody extraction). PianoBART [22] extends this with BART-style denoising and multi-level masking over Octuple-encoded piano MIDI, unifying understanding and generation in a single encoder-decoder.

Autoregressive foundation models: A parallel line of work treats MIDI as a language and trains large causal transformers. ChatMusician [23] and MuseCoco [24] target controllable composition from text, while Moonbeam [25] scales next-token prediction to 18 B MIDI tokens, pre-training on ∼81,600 hours of data using a Fundamental Music Embedding tokenizer and multidimensional relative attention. Aria [26] focuses on solo piano, pre-training on ∼60,000 hours of MIDI and obtaining strong embeddings via subsequent contrastive fine-tuning.

Contrastive and cross-modal alignment: CLaMP 3 [27] aligns multilingual text, symbolic music (ABC and MIDI), and audio in a shared embedding space through contrastive learning, supporting retrieval across modality pairs that were never jointly observed during training. In the piano-performance setting, Pianist Transformer [28] removes the need for score-performance alignment and pre-trains on large unaligned corpora for expressive rendering.

Downstream evaluation of these representations has centred on composition-level classification (genre, emotion, chord recognition). Performance-level understanding, in contrast, remains fragmented across isolated studies, including difficulty estimation [29], pianist identification [30], and score-informed mistake detection [31]. Building on these downstream evaluation tasks and inspired by the Evaluation Package for Audio Representations (*EVAR*), the unified EVPMR benchmark is introduced in this work.

2.2 Joint-Embedding Predictive Architecture (JEPA)

JEPA [15] learns representations by predicting target latent embeddings from partial context without reconstruction or contrastive objectives. Using an asymmetric encoder–predictor design, it focuses on modelling predictable structure in latent space, enabling abstract and semantically meaningful representations across domains. This framework has been used to learn semantic image representations

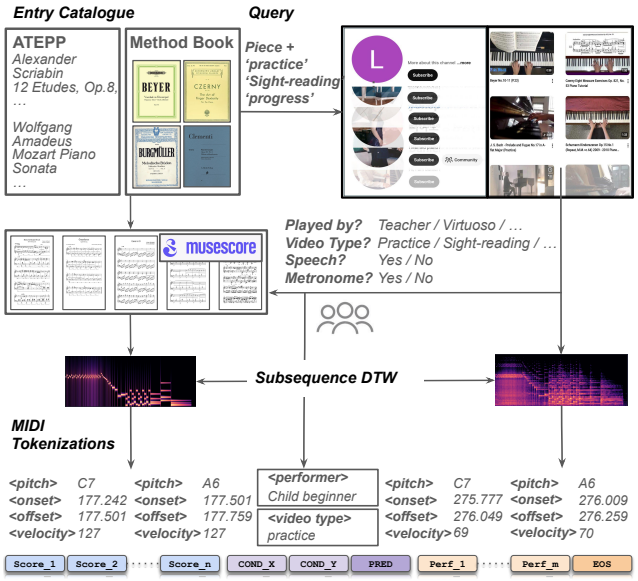

Figure 1. Data curation pipeline. YouTube recordings are collected from entry catalogues and search queries, annotated with performer expertise and recording context, aligned with the corresponding score via subsequence DTW. Aligned pairs are segmented and tokenised for pre-training.

by predicting masked spatial target blocks from context (I-JEPA [32]), and spatio-temporal representations by predicting latent targets from video context (V-JEPA 2 [33]). Audio-JEPA [34] translates the JEPA principle to the audio domain by predicting latent representations of masked mel-spectrogram patches.

In the music domain, Stem-JEPA [35] leverages the multi-track structure of music to replace input-space masking with stem holdout, training the predictor to predict the latent representation of a missing instrument from a context mix conditioned on instrument labels. JEP-AGA [36] combines the principles of JEPA with autoregressive generation for music infilling. AMT-JEPA [37] demonstrates that JEPA can learn semantically meaningful representations for multi-instrument music transcription, particularly capturing instrument and pitch-related information. Hachana and Rasheed [38] extend JEPA to symbolic domain with music-specific masking strategies for MIDI sequences.

3. DATA CURATION

Existing symbolic music datasets focus predominantly on professional performances: ATEPP [13] contains virtuoso interpretations and ASAP [14] provides competition-level recordings with aligned scores. To address this gap, we curate the **MAJEPPA dataset**, comprising 3,979 real-world piano recordings across 942 distinct pieces and movements, spanning the full performance spectrum from child beginners to concert virtuosi (Figure 2).

Collection: Solo-piano recordings that are publicly available are collected from YouTube, covering demo recordings, informal practice, and sight-reading sessions. To ensure diversity of performance levels across the music pieces,

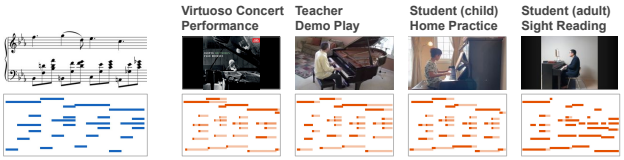

Figure 2. The same piece realised across different expertise levels and recording contexts. Left: score excerpt with MIDI piano roll. Right: four performances showing visible differences in timing, note accuracy, and dynamics.

we source our data from catalogues and channels (top of Figure 1). Firstly, we query for repertoire titles, adding keywords such as *Practice, Sight-reading*, or *Progress*. For concert-repertoire level pieces, we query for videos of the same pieces or movements as those in ATEPP. To include beginner-level material, we query for exercises and pedagogical repertoire commonly used in piano training, such as method books by Carl Czerny, Friedrich Burgmüller, and Anton Diabelli. We also include widely practised contemporary pieces such as *Kiss the Rain* by Yiruma. From the results, we manually identify resourceful channels from piano teachers (demonstrations and tutorials), sight-reading sessions, practice recordings, and learning progress videos.

3.1 Annotation

After sourcing videos, each recording is manually annotated by conservatoire student pianists [1] who work from a shared calibration document that defines each category and provides reference clips. The annotations capture differences along two orthogonal axes: recording context and performer expertise (Table 1). Recording context describes the setting in which the performance was captured:

- **Practice**: Recordings from students who are actively learning a particular piece.
- **Sight-reading**: Sight-reading sessions involving one or multiple unfamiliar pieces.
- **Showcase**: Recordings presented to teachers (e.g., prior to weekly lessons) or in student concerts, reflecting a more polished, near-standard performance with relatively fewer mistakes despite the performer's developing skill level.
- **Demo**: Demonstration recordings by piano teachers illustrating how a piece should be performed. They are usually highly accurate in terms of pitch and timing, and not overly expressive.
- **Slow demo**: Slow-tempo demonstration recordings by piano teachers.
- **Concert**: Commercial recordings or live performances from pianists.

Performer expertise describes the background of the performer:

- **Child beginner**: Young learners performing simple exercises or early-stage repertoire.

[1] Annotator reliability: `https://jepa-demo.vercel.app`

Table 1. Distribution of recordings by performer expertise (rows) and recording context (columns).

	Prac.	Sight.	Show.	Demo	Slow	Conc.	**Total**	**Hours**
Teacher	35	15	55	690	311	173	1279	76.1
Ad. Beg.	622	217	92	12	3	1	947	127.9
Ad. Int.	444	135	177	16	2	3	777	50.7
Ch. Beg.	186	8	193	2	2	0	391	8.9
Ch. Pro.	84	0	161	2	0	0	247	13.7
Virtuoso	0	0	0	0	0	338	338	28.9
Total	1371	375	678	722	318	515	**3979**	303.2

- **Child professional**: Young performers demonstrating advanced technical proficiency, typically on a pre-professional or professional training track.
- **Adult beginner**: Adult learners performing introductory exercises or beginner repertoire, typically engaging with piano as a hobby.
- **Adult intermediate**: Performers demonstrating moderate technical proficiency and developing musicality.
- **Piano teacher**: Instructors providing demonstration performances for instructional purposes.
- **Virtuoso**: Established professional pianists.

Besides the conditioning tags, the annotators also provide the additional information: **1). Breakpoints** for long videos, especially when multiple pieces are in a single recording; **2). Corresponding scores** sourced from *MuseScore* for each identified piece, with annotators visually matching the score edition to the performed arrangement (e.g., key, structure, and approximate length); **3). Filtering out** videos that contain speech, metronome, or other non-piano sounds, and those not meeting the dataset criteria.

3.2 Processing

The final processing pipeline produces aligned pairs of score and performance MIDI segments: **1). Breakpoint processing**: long recordings are first segmented into individual performance excerpts based on annotation. Each segment is associated with its corresponding score; **2). Transcription**: performance audio is transcribed into MIDI using Transkun [39], chosen as the note-level state of the art on standard piano AMT benchmarks rather than alternatives such as Aria-AMT [40], though we did not explicitly check transcription quality (including pedal artefacts); **3). Alignment**: to enable segment-level pairing under the limited context length of language models, score and performance must be temporally aligned before cropping. Due to the highly non-ideal and fuzzy nature of real-world data (e.g., mistakes, omitted passages, one-hand practice), we perform subsequence dynamic time warping (DTW) on chroma features extracted from synthesised score audio and performance audio, allowing robust partial alignment. Subsequence DTW handles omitted repeats by aligning the performance to the relevant contiguous score portion; performances with extra repetitions may produce locally noisier pairs, which the segment-level objective tolerates. The resulting warping path maps score time to performance time, enabling consistent segmentation into fixed-token-length windows for

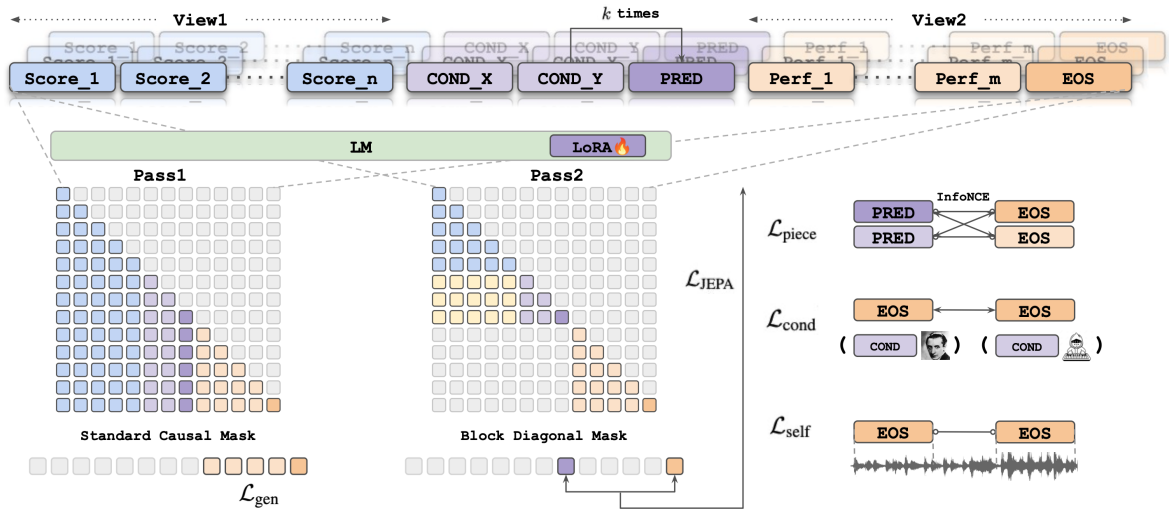

Figure 3. Sequence packing and attention mask design. **Left:** standard causal mask for the generative item. **Middle:** block-diagonal causal mask for the JEPA item, enforcing independent encoding of score and performance views. **Right:** contrastive objectives for piece alignment, cross-condition clustering, and within-performance alignment.

downstream processing.

4. PROPOSED METHOD

Our goal is to learn a unified representation of piano performance that captures both the underlying musical content and its expressive realisation. To this end, we model score and performance as two complementary views of the same musical excerpt, and learn to predict performance representations from score representations in a shared latent space.

We formulate each training example as a *score–performance pair* $(\mathbf{s}, \mathbf{p})$ corresponding to the same musical excerpt. The score view $\mathbf{s}$ is a deadpan tokenisation of the written score, while the performance view $\mathbf{p}$ is the tokenised expressive realisation of the same music. Both share the same musical content, but differ in the performance detail: the score specifies the intended notes and structure, whereas the performance carries timing, dynamics, articulation, and recording context. Following LLM-JEPA [16], we treat this pair as a natural two-view structure and train a predictor to map from the score view to the performance view in representation space. A mini-batch contains B such pairs.

The objective is to learn a unified representation of piano performance that supports both conditional generation and abstract embedding prediction. Specifically, the model is trained to (i) generate performance tokens conditioned on score tokens, and (ii) predict a latent representation of the performance from the score and conditioning tokens alone.

4.1 Model components

We implement this two-view predictive framework within a single autoregressive transformer by augmenting it with specialised tokens and projection heads. The architecture comprises an encoder backbone, a predictor via `[PRED]` tokens, conditioning tokens encoding performer and recording context, and a projection head for contrastive objectives.

Encoder. We use Aria-medium [26], a 632M-parameter LLaMA-style autoregressive transformer pre-trained on symbolic MIDI. **Predictor.** We append k `[PRED]` tokens to the input sequence and reuse the model's self-attention as a tied-weights predictor: the final `[PRED]` state represents $\mathrm{Pred}(\mathbf{s})$. **Condition tokens.** Two tokens, `[COND_X]` (performer) and `[COND_Y]` (recording context), provide categorical conditioning via learned embeddings (Section 3).

4.2 Sequence construction and attention masking

Each score–performance pair $(\mathbf{s}, \mathbf{p})$ is first packed into a single token sequence as shown in Fig. 3: $[\mathbf{s};\ \texttt{COND};\ \texttt{PRED};\ \mathbf{p};\ \texttt{EOS}]$, where $\mathbf{s}$ denotes the score tokens, $\mathbf{p}$ the performance tokens, and `COND`, `PRED`, and `EOS` are special tokens marking the conditioning boundary, prediction query, and sequence termination, respectively.

Pass 1 – generative (causal mask). The packed sequence $[\mathbf{s};\ \texttt{COND};\ \texttt{PRED};\ \mathbf{p};\ \texttt{EOS}]$ is processed under a standard lower-triangular causal mask, so performance tokens attend to the full preceding context. This pass produces the next-token logits over the performance tokens used by the generative loss $\mathcal{L}_{\text{gen}}$ defined below.

Pass 2 – JEPA (block-diagonal causal mask). The same tokens are processed under a block-diagonal mask that defines two independent views: View 1 $[\mathbf{s};\ \texttt{COND};\ \texttt{PRED}]$ and View 2 $[\mathbf{p};\ \texttt{EOS}]$, each internally causal but mutually invisible. We extract $\mathrm{Pred}(\mathbf{s})$ from the `[PRED]` position in View 1 and $\mathrm{Enc}(\mathbf{p})$ from the `[EOS]` position in View 2. By construction (Fig. 3), `[PRED]` attends only to score and conditioning tokens.

This split ensures (i) Pass 1 retains the model's ability to generate performances conditioned on scores, and (ii) Pass 2 must form an abstract prediction from the score alone, since `[PRED]` never observes the performance tokens.

4.3 Training objective

The training loss combines a generative term with contrastive terms that align score and performance representations, enforce condition-aware structure, and maintain

within-performance consistency. For each pair i in the batch, we obtain two embeddings via a shared linear projection head $\phi : \mathbb{R}^{1536} \to \mathbb{R}^{512}$, both ℓ_2-normalised:

$$\mathbf{z}_i^s = \phi(\mathrm{Pred}(\mathbf{s}_i)) \quad \text{and} \quad \mathbf{z}_i^p = \phi(\mathrm{Enc}(\mathbf{p}_i)),$$

where $\mathrm{Pred}(\mathbf{s}_i)$ and $\mathrm{Enc}(\mathbf{p}_i)$ are the Pass 2 outputs from the `[PRED]` and `[EOS]` positions respectively. The training loss combines a generative term computed from Pass 1 with three contrastive JEPA terms computed from Pass 2:

$$\mathcal{L} \;=\; \gamma\,\mathcal{L}_{\text{gen}} \;+\; \lambda\left(\mathcal{L}_{\text{piece}} + \beta\,\mathcal{L}_{\text{cond}} + \alpha\,\mathcal{L}_{\text{self}}\right), \quad (1)$$

where the four loss terms are defined below and $\gamma, \lambda, \beta, \alpha \geq 0$ are scalar weights. All contrastive terms share a single temperature τ.

Generative loss $\mathcal{L}_{\text{gen}}$ is the standard next-token cross-entropy over the performance tokens of $\mathbf{p}$, conditioned on score and conditioning tokens through the Pass 1 causal mask.

Piece alignment $\mathcal{L}_{\text{piece}}$ is an InfoNCE loss that pulls each predicted score embedding towards its paired performance embedding within the batch:

$$\mathcal{L}_{\text{piece}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\mathbf{z}_i^s\cdot\mathbf{z}_i^p/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{z}_i^s\cdot\mathbf{z}_j^p/\tau)}. \quad (2)$$

Cross-condition clustering $\mathcal{L}_{\text{cond}}$ is a soft supervised contrastive loss over performance embeddings. Let $w_{ij} \in [0,1]$ denote a continuous similarity between samples i and j, derived from the ordinal distance between their performer skill levels and recording formality levels, and let $\mathcal{C}_i = \{j : \text{cond}(j) \neq \text{cond}(i)\}$ restrict to pairs with different condition tags:

$$\mathcal{L}_{\text{cond}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j\in\mathcal{C}_i}\bar{w}_{ij}\log\frac{\exp(\mathbf{z}_i^p\cdot\mathbf{z}_j^p/\tau)}{\sum_{h\in\mathcal{C}_i}\exp(\mathbf{z}_i^p\cdot\mathbf{z}_h^p/\tau)}, \quad (3)$$

with $\bar{w}_{ij} = w_{ij}/\sum_{h\in\mathcal{C}_i} w_{ih}$ row-normalised over $\mathcal{C}_i$. This term replaces the cosine-distance loss of [16], which we found prone to representation collapse in our setting. See our website [1] for details of continuous condition similarity design.

Within-performance alignment $\mathcal{L}_{\text{self}}$ is a symmetric InfoNCE loss over pairs of segments from the same performance. We form B pairs $(\mathbf{p}_i^a, \mathbf{p}_i^b)$ with embeddings $\mathbf{z}_i^a, \mathbf{z}_i^b$, and define

$$\mathcal{L}_{\text{self}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{u\in\{a,b\}}\log\frac{\exp(\mathbf{z}_i^u\cdot\mathbf{z}_i^{\bar{u}}/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{z}_i^u\cdot\mathbf{z}_j^{\bar{u}}/\tau)}, \quad (4)$$

where $\bar{a} = b$ and $\bar{b} = a$. Both terms encourage segments from the same performance to be more similar than segments from different recordings, while averaging over the two directions yields a symmetric objective. Since the two segments share performer and recording tags, $\mathcal{L}_{\text{self}}$ complements $\mathcal{L}_{\text{cond}}$ by encouraging within-performance coherence.

Training setup: We freeze the pre-trained Aria weights and apply Low-Rank Adaptation (LoRA) [41] to the attention projections `mixed_qkv` and `att_proj_linear` (rank $r = 512$, scaling $\alpha_{\text{LoRA}} = 32$), giving ∼75M trainable parameters out of 660M. The context window is extended from 2048 to 4096 tokens via linear RoPE scaling [42]. Loss hyperparameters in Eq. (1) are set to $\gamma = 1$, $\lambda = 8$, $\beta = 1$, $\alpha = 1$, with temperature $\tau = 0.07$ and $k = 1$ `[PRED]` token, chosen by ablation.

We train on 4×NVIDIA A100-80GB GPUs in bfloat16 mixed precision for 120 epochs, with a per-GPU batch size of 8 and gradient accumulation of 4. Optimisation uses AdamW (weight decay 0.01) with separate learning rates for these LoRA modules (2×10^{-4}), the new conditioning and `[PRED]` token embeddings (1×10^{-3}), and the LM head (1×10^{-5}).

5. DOWNSTREAM TASKS AND EVALUATION

Inspired by EVAR, we formulate performance-related tasks in a unified package, EVPMR, providing standardised evaluation protocols. All audio-only datasets are transcribed to MIDI with Transkun [39] for consistency. Every task is evaluated with 5-fold cross-validation repeated across 5 random seeds. On top of a frozen global embedding from `encode()`, classification uses a linear probe, pairwise ranking uses a binary classification over concatenated pair features, regression uses Ridge, and UMP uses a frame-level MLP decoder.

Performance assessment. Three datasets provide normalised ratings in $[0, 1]$: NeuroPiano [4] (104 recordings, 27 students, 6 pieces, rated by 33 teachers on a 0–6 scale), YCU-PPE-III [43] (606 performances, 13 songs, rated by 3 instructors on a 0–100 scale), and PISA [44] (59 YouTube videos, skill level 1–10). Ratings are aggregated per recording and normalised using min–max scaling. We report MAE↓ as the primary metric.

Chopin competition ranking. Following Piano-Judge [5], this task uses 130 preliminary-round performances from the 2015 International Chopin Piano Competition, each labelled with the round the performer reached (0–4). We cast it as pairwise ranking: for every ordered pair of performances with different labels, the probe predicts which of the two advanced further. We report pairwise ranking accuracy (chance = 0.50).

Technique detection. From the Pianism-Labelling Dataset [5] we use the single-label piano technique classification task, which assigns each performance excerpt to one of 7 technique categories (octaves, scales, etc.).

Pianist8. Introduced with MidiBERT-Piano [21], this dataset contains 411 pieces by 8 pianists spanning four stylistic categories. The task is 8-way pianist identification, testing whether representations capture individual styles.

Unreferenced Mistake Prediction (UMP). Following the score-free "conspicuous mistake" formulation introduced by [1], UMP detects performance errors from musical context alone. We evaluate on three open-source datasets: EN-augmented (*EN*) [45], Burgmüller (*BM*) [1], and ScoreInformed (*SI*) [46]. UMP is token-level: we extract per-token embeddings, assign each token a timestamp and duration, and group them into a fixed 30 ms grid (duration-aware: each token contributes to all frames it

Table 2. Downstream evaluation on the EVPMR benchmark. Mean (std.) over 5-fold cross-validation across 5 random seeds.

	Chopin	Technique	Quality Regression: MAE↓			UMP: $F1$			Pianist8
Model	PairAcc	Acc	PISA	NeuroPiano	YCU-PPE	EN	BM	SI	Acc
Aria	47.5 (8.1)	66.2 (11.4)	.122 (.031)	.215 (.019)	.068 (.005)	.174 (.025)	.313 (.016)	.300 (.020)	86.1 (4.3)
Moonbeam	53.7 (2.9)	77.9 (4.8)	.151 (.035)	.192 (.014)	.061 (.004)	–	–	–	80.6 (4.5)
MusicBERT	52.5 (5.8)	67.1 (4.6)	.190 (.026)	.205 (.019)	.065 (.004)	–	–	–	78.6 (2.1)
CLaMP 3	53.5 (8.5)	70.3 (2.7)	.217 (.036)	.203 (.017)	.089 (.004)	–	–	–	80.8 (3.3)
MAJEPPA ($\mathcal{L}_{\text{piece}}$)	53.7 (8.1)	76.3 (5.9)	.125 (.083)	.180 (.015)	.064 (.004)	.151 (.021)	.308 (.012)	.313 (.015)	83.1 (3.2)
MAJEPPA ($\mathcal{L}_{\text{piece}}$+$\mathcal{L}_{\text{cond}}$)	58.7 (8.7)	70.3 (9.5)	.108 (.090)	.181 (.022)	.062 (.005)	.141 (.015)	.321 (.012)	**.341** (.004)	**86.8** (5.2)
MAJEPPA ($\mathcal{L}_{\text{piece}}$+$\mathcal{L}_{\text{cond}}$+$\mathcal{L}_{\text{self}}$)	**59.7** (9.4)	**79.0** (6.5)	**.090** (.034)	**.155** (.018)	**.059** (.002)	**.179** (.012)	**.333** (.015)	.316 (.037)	84.2 (3.8)

spans). A per-frame MLP probe predicts binary mistake labels; predictions are merged into segments and evaluated with `mir_eval.transcription` (segment $F1$, 2 s tolerance). As this protocol relies on the model's tokenisation and onset/duration information, we therefore compare MAJEPPA only against the Aria baseline on UMP as other baselines use incomparable token schemes.

Results: As reported in Table 2, the full MAJEPPA model attains the best score on Chopin, Technique, all regression tasks, and UMP EN/BM, while the $\mathcal{L}_{\text{piece}}$+$\mathcal{L}_{\text{cond}}$ variant is best on Pianist8 and UMP SI. MAJEPPA variants exceeds four baselines on Chopin, Technique, and the regression tasks, indicating that score-conditioned JEPA pre-training captures expressive and pedagogical structure not recovered by purely generative (Aria, Moonbeam) or masked-reconstruction (MusicBERT, CLaMP 3) pre-training. Aria's tokenizer has no explicit pedal tokens (sustain is folded into note durations), so the Aria baseline comparison does not isolate pedalling as a discrete cue.

The ablation reveals a division of labour among the three losses. $\mathcal{L}_{\text{piece}}$ alone is already competitive with the strongest baselines on most tasks. Adding $\mathcal{L}_{\text{cond}}$ benefits the tag-discriminative tasks (Pianist8, UMP SI), where samples with different but similar condition tags are brought closer according to w_{ij}. Adding $\mathcal{L}_{\text{self}}$ further reduces regression MAE and improves Chopin and UMP EN/BM, but slightly degrades Pianist8 and UMP SI: pulling within-performance segments together smooths over the segment-level cues these two tasks rely on.

5.1 Generation and embedding space analysis

Due to time constraints, we did not conduct a full-scale evaluation of the generation capabilities. However, we invite readers to visit our website [1] for sample outputs and interactive generation. In our demo website, we also showcase representative trajectories of PCA-reduced embeddings.

Practice progression. On the session-averaged embeddings of 7 YouTube practice journeys, we measure Spearman rank correlations between session index (from earliest to latest session) and the trajectory's leading principal component. The mean correlation is $\rho = 0.41$ across PC1 and PC2, with trajectories exhibiting strong monotonic drift. Even in cases with moderate correlation, the latest-session points consistently concentrate at one end of each trajectory cloud (Figure 4), suggesting that MAJEPPA induces a directional notion of progress without explicit temporal supervision.

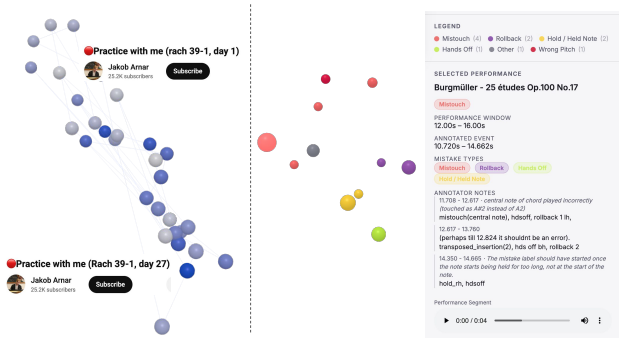

Figure 4. Samples from embedding viewer. **Left**: Practice trajectories for 30 days of Rachmaninov practice, coloured white (early) to blue (late). **Right**: UMAP of mistake regions for Burgmüller Étude Op. 100 No. 17 with annotations coloured by canonical mistake type, matching the interactive website.

Mistake topology. Across the 25 Burgmüller études, we project mistake-annotated 4-second windows to 3D via UMAP, and further label 94 of the annotated regions with categorical mistake types (mistouch, wrong-pitch, rollback, ...) drawn from independent human annotations in the SynMist project [2]. Colouring by mistake type (Figure 4, right) reveals contiguous within-piece subregions: timing-disruption mistakes (rollback / hands-off / hold) sit apart from pitch-deviation mistakes (mistouch / wrong-pitch), suggesting that MAJEPPA encodes the texture of a deviation rather than a single "is-a-mistake" axis.

6. CONCLUSION

We presented MAJEPPA, a dataset and framework for unified piano performance representations. We contribute ∼4,000 annotated recordings across six expertise levels and six recording contexts, spanning from child beginners to concert virtuosi. We use the dataset to train an LLM-JEPA model for symbolic music using a joint objective combining score-conditioned generation, cross-condition clustering, and within-performance alignment. MAJEPPA establishes a new representation space for music education, where performance quality, progression, and technique can be modelled, compared, and guided in a unified and data-driven manner.

7. ETHICS STATEMENT

MAJEPPA is built from publicly available YouTube solo-piano recordings for research purposes; because these recordings are publicly accessible, we did not obtain individual consent from uploaders. The corpus includes children's practice and showcase videos and commercial or concert recordings by professional pianists. However, public availability does not constitute a waiver of the uploaders' or performers' underlying rights. To mitigate this, we release only derived MIDI and coarse descriptive tags (expertise and recording context), not raw audio or personally identifying metadata beyond what is already public, and we will respond promptly to any request from an uploader, performer, or rights holder to remove specific content from the dataset. Additionally, we intend the resource for research on performance understanding and music education rather than for replacing teachers or performers.

8. ACKNOWLEDGMENTS

This work is supported by the UKRI Centre for Doctoral Training in Artificial Intelligence and Music. J. Zhou is a research student supported jointly by the China Scholarship Council and Queen Mary University of London.

9. REFERENCES

[1] A. Morsi, K. Tatsumi, A. Maezawa, T. Fujishima, and X. Serra, "Sounds out of pläce? Score-Independent detection of conspicuous mistakes in piano performances," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2023, pp. 352–358.

[2] A. Morsi, H. Zhang, A. Maezawa, S. Dixon, and X. Serra, "Simulating piano performance mistakes for music learning," in *Proceedings of the 21st Sound and Music Computing Conference (SMC)*, 2024, pp. 179–186.

[3] S. Chang, S. Dixon, and E. Benetos, "RUMAA: Repeat-aware unified music audio analysis for score-performance alignment, transcription, and mistake detection," in *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025, pp. 1–5.

[4] H. Zhang, V. Cheung, H. Nishioka, S. Dixon, and S. Furuya, "How does the teacher rate? Observations from the NeuroPiano dataset," in *International Society for Music Information Retrieval (ISMIR) Late Breaking Demo (LBD)*, 2024.

[5] H. Zhang, J. Liang, and S. Dixon, "From audio encoders to piano judges: Benchmarking performance understanding for solo piano," in *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024, pp. 511–519.

[6] H. Zhang, V. K. Cheung, H. Nishioka, S. Dixon, and S. Furuya, "LLaQo: Towards a query-based coach in expressive music performance assessment," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[7] Š. Libřický and J. Hajič, "Modeling the difficulty of saxophone music," in *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, 2025, pp. 747–754.

[8] M. Müllerschon, A. Klapuri, M. Rodriguez, and C. Cardin, "Playability prediction in digital guitar learning using interpretable student and song representations," in *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, 2025, pp. 631–637.

[9] Z. Hassein-Bey, Y. Abbou, A. D'Hooge, M. Giraud, G. Guillemain, and A. Jeanneau, "What song now? Personalized rhythm guitar learning in western popular music," in *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, 2025, pp. 296–302.

[10] A. Morsi, S. Chiruthapudi, S. D. Peter, M. Pilkov, L. Bishop, A. Maezawa, X. Serra, and C. E. Cancino-Chacón, "Enabling empirical analysis of piano performance rehearsal with the Rach3 MIDI dataset," in *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, 2025, pp. 484–491.

[11] Y. Kim, J. Park, J. Bae, K. Kim, T. Kwon, A. Lerch, and J. Nam, "PianoVAM: A multimodal piano performance dataset," in *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, 2025, pp. 528–535.

[12] J. Choi, T. Kwon, and J. Nam, "Predicting flutist onset timing in duet performance: A multimodal analysis of gesture and breath cues," in *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, 2025, pp. 100–106.

[13] H. Zhang, J. Tang, S. R. Rafee, S. Dixon, G. Fazekas, and G. A. Wiggins, "ATEPP: A dataset of automatically transcribed expressive piano performance," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru,India, 2022, pp. 446–453.

[14] F. Foscarin, A. McLeod, P. Rigaux, F. Jacquemard, and M. Sakai, "ASAP: A dataset of aligned scores and performances for piano transcription," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020, pp. 534–541.

[15] Y. LeCun, "A path towards autonomous machine intelligence," *Open Review*, vol. 62, no. 1, pp. 1–62, 2022.

[16] H. Huang, Y. LeCun, and R. Balestriero, "LLM-JEPA: Large language models meet joint embedding predictive

architectures," in *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026.

[17] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20, 2020, pp. 1180–1188.

[18] W.-Y. Hsiao, J.-Y. Liu, and Y.-H. Yang, "Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, 2021, pp. 178–186.

[19] M. Zeng, X. Tan, R. Wang, Z. Ju, T. Qin, and T.-Y. Liu, "MusicBERT: Symbolic music understanding with large-scale pre-training," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 791–800.

[20] J. Lenz and A. Mani, "PerTok: Expressive encoding and modeling of symbolic musical ideas and variations," in *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024, pp. 981–988.

[21] Y.-H. Chou, I.-C. Chen, C.-J. Chang, J. Ching, and Y.-H. Yang, "MidiBERT-Piano: Large-scale pre-training for symbolic music understanding," *Journal of Creative Music Systems (JCMS)*, vol. 8, no. 1, 2024.

[22] X. Liang, Z. Zhao, W. Zeng, Y. He, F. He, Y. Wang, and C. Gao, "PianoBART: Symbolic piano music generation and understanding with large-scale pre-training," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 2024, pp. 1–6.

[23] R. Yuan, H. Lin, Y. Wang, Z. Tian, S. Wu, T. Shen, G. Zhang, Y. Wu, C. Liu, Z. Zhou, Z. Ma, L. Xue, Z. Wang, Q. Liu, T. Zheng, Y. Li, Y. Ma, Y. Liang, X. Chi, R. Liu, Z. Wang, P. Li, J. Wu, C. Lin, Q. Liu, T. Jiang, W. Huang, W. Chen, E. Benetos, J. Fu, G. Xia, R. Dannenberg, W. Xue, S. Kang, and Y. Guo, "ChatMusician: Understanding and generating music intrinsically with LLM," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 6252–6271.

[24] P. Lu, X. Xu, C. Kang, B. Yu, C. Xing, X. Tan, and J. Bian, "MuseCoco: Generating symbolic music from text," *arXiv preprint arXiv:2306.00110*, 2023.

[25] Z. Guo and S. Dixon, "Moonbeam: A midi foundation model using both absolute and relative music attributes," *arXiv preprint arXiv:2505.15559*, 2025.

[26] L. Bradshaw, H. Fan, A. Spangher, S. Biderman, and S. Colton, "Scaling self-supervised representation learning for symbolic piano performance," in *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, 2025, pp. 451–459.

[27] S. Wu, Z. Guo, R. Yuan, J. Jiang, S. Doh, G. Xia, J. Nam, X. Li, F. Yu, and M. Sun, "CLaMP 3: Universal music information retrieval across unaligned modalities and unseen languages," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 2605–2625.

[28] H.-J. You, J.-J. Shao, X.-W. Yang, L.-H. Jia, L.-Z. Guo, and Y.-F. Li, "Pianist transformer: Towards expressive piano performance rendering via scalable self-supervised pre-training," *arXiv preprint arXiv:2512.02652*, 2025.

[29] P. Ramoneda, D. Jeong, V. Eremenko, N. C. Tamer, M. Miron, and X. Serra, "Combining piano performance dimensions for score difficulty classification," *Expert Systems with Applications*, p. 121776, 2024.

[30] J. Tang, G. Wiggins, and G. Fazekas, "Pianist identification using convolutional neural networks," in *Proceedings of the 4th International Symposium on the Internet of Sounds (IS2)*, 2023.

[31] B. S.-H. Chou, P. Jajal, N. J. Eliopoulos, J. C. Davis, G. K. Thiruvathukal, K. Y.-J. Yun, and Y.-H. Lu, "Laddersym: A multimodal interleaved transformer for music practice error detection," in *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

[32] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 15 619–15 629.

[33] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, M. Komeili, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus, S. Arnaud, A. Gejji, A. Martin, F. R. Hogan, D. Dugas, P. Bojanowski, V. Khalidov, P. Labatut, F. Massa, M. Szafraniec, K. Krishnakumar, Y. Li, X. Ma, S. Chandar, F. Meier, Y. LeCun, M. Rabbat, and N. Ballas, "V-JEPA 2: Self-supervised video models enable understanding, prediction and planning," *arXiv preprint arXiv:2506.09985*, 2025.

[34] L. Tuncay, E. Labbé, E. Benetos, and T. Pellegrini, "Audio-JEPA: Joint-embedding predictive architecture for audio representation learning," in *ICME 2025*, Nantes, France, 2025.

[35] A. Riou, S. Lattner, G. Hadjeres, M. Anslow, and G. Peeters, "Stem-JEPA: A joint-embedding predictive architecture for musical stem compatibility estimation," in *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[36] Z. Hu, Y. Liu, G. Chen, and B. X. B. Yu, "Compose with me: Collaborative music inpainter for symbolic music infilling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 2, 2025, pp. 1327–1335.

[37] M. Pilataki, M. Mauch, and S. Dixon, "Self-supervised representation learning with a JEPA framework for multi-instrument music transcription," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025, pp. 1–5.

[38] R. Hachana and B. Rasheed, "Using a joint-embedding predictive architecture for symbolic music understanding," in *AI for Music Workshop*, 2025.

[39] Y. Yan and Z. Duan, "Scoring time intervals using non-hierarchical transformer for automatic piano transcription," in *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024, pp. 973–980.

[40] L. Bradshaw and EleutherAI, "Aria-amt: Efficient and robust automatic piano transcription," 2024. [Online]. Available: https://github.com/EleutherAI/aria-amt

[41] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[42] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[43] W. Wang, J. Pan, H. Yi, Z. Song, and M. Li, "Audio-based piano performance evaluation for beginners with convolutional neural network and attention mechanism," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 1119–1133, 2021.

[44] P. Parmar, J. Reddy, and B. Morris, "Piano skills assessment," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, 2021, pp. 1–5.

[45] Y. Jiang, "Expert and novice evaluations of piano performances: Criteria for computer-aided feedback," in *Proceedings of the 24th International Society for Music Information Retrieval (ISMIR)*, 2023, pp. 367–374.

[46] E. Benetos, A. Klapuri, and S. Dixon, "Score-informed transcription for automatic piano tutoring," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 2153–2157.