超越文件级划分：通过多视图对齐重新思考符号音乐评估

Kun Bai、Yinghan Wang、Canran Xiao、Jiachen Luo、Huy Phan、Yong Dai、Jiaqi Zhang、Yunwei Li、Lin Wang、Joshua Reiss 和 Björn W. Schuller，*IEEE* 会士

arXiv:2611.28765v1 [cs.SD] 21 Nov 2026

摘要—符号音乐生成通常通过预测后续音符或事件来续写 MIDI 乐段。当训练数据中包含测试作品的另一种演奏版本或编曲时，留出预测损失可能会高估模型对未见作品的泛化能力。我们提出了多视图对齐感知同作品识别方法，将互补的音乐表示与共享乐段的时间对齐相结合，从训练集中剔除测试作品的不同版本，同时保留无关内容并固定评估文件。在一个受控样本组中，我们选取了 100 对经人工核验且嵌入相似度较高的样本。用测试作品版本替换 3.29% 的训练窗口后，相较于匹配的无关内容，各组平均负对数似然最多降低 30.16%。在两个数据集和两种生成器上，我们的方法使测试损失最接近基于独立作品标签构建的参考值。与随机移除相同数量的文件相比，该方法减少了暴露的测试作品数量，并保留了更多符合条件的内容；在钢琴演奏数据集上，保留了约 99% 的符合条件词元。针对该数据集的另一项研究匹配了训练词元数量和更新步数：根据数据集标签，没有测试作品仍处于暴露状态，平均相对测试损失比独立选取的参考值高 2.15%。这些发现支持在评估模型对未见作品的泛化能力时，将同一作品的所有版本置于同一数据划分中。项目网址为 https://kunbai05-nb.github.io/。

一、引言

SYMBOLIC 用于作品创作、续写和伴奏的音乐生成模型 [1],

Kun Bai 任职于北京建筑大学智能科学与技术学院，地址：中国北京 102616（电子邮箱：kunbai05@gmail.com）。

Yinghan Wang 任职于北京林业大学理学院，地址：中国北京 100083（电子邮箱：yinghan@bjfu.edu.cn）。

Canran Xiao 任职于中山大学，地址：中国深圳 518107（电子邮箱：xiaocr3@mail.sysu.edu.cn）。

Jiachen Luo 任职于慕尼黑工业大学计算、信息与技术学院，地址：德国慕尼黑 80333；同时还任职于伦敦玛丽女王大学电子工程与计算机科学学院数字音乐中心，英国伦敦 E1 4NS（电子邮箱：jiachen.luo@qmul.ac.uk）。

Huy Phan 就职于德国人工智能研究中心（DFKI），德国吕贝克 23562（电子邮箱：huy.phan@dfki.de）。

Yong Dai 就职于 X-Humanoid，中国北京 101111（电子邮箱：daiyongya@outlook.com）。

Jiaqi Zhang 任职于中国江苏大学电子与通信工程学院，地址为中国镇江 212013（电子邮箱：3230602065@stmail.ujs.edu.cn）。

Yunwei Li 任职于美国华盛顿州西雅图市华盛顿大学，邮编 98195（电子邮箱：yunweili@uw.edu）。

Lin Wang 和 Joshua Reiss 任职于英国伦敦玛丽女王大学电子工程与计算机科学学院数字音乐中心，地址为伦敦 E1 4NS（电子邮箱：{lin.wang,joshua.reiss}@qmul.ac.uk）。

Björn W. Schuller 任职于德国慕尼黑 81675 的慕尼黑工业大学健康信息学教席，同时任职于英国伦敦 SW7 2AZ 的伦敦帝国理工学院语言、音频与音乐研究组（电子邮件：schuller@tum.de）。

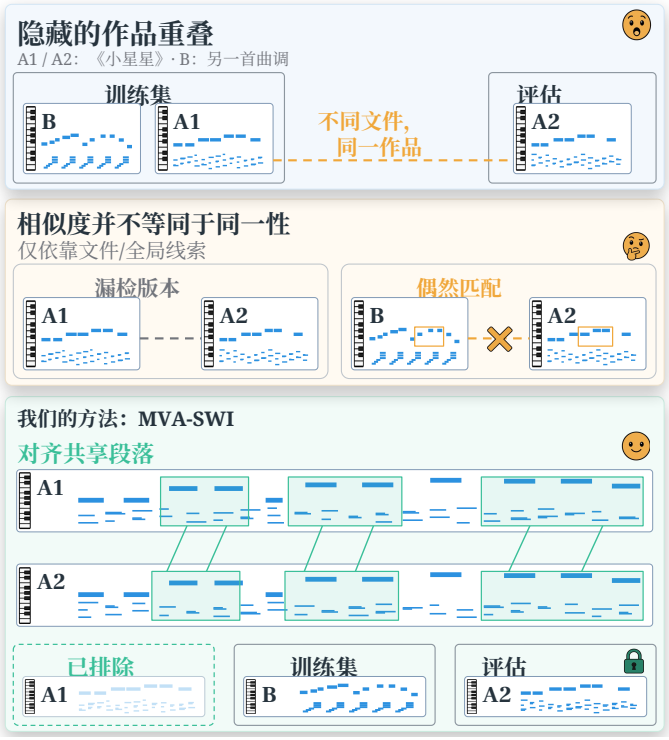


图 1. 作品级隔离示例。A1 和 A2 是《一闪一闪小星星》的不同演奏版本，而 B 是另一首曲调。MVA-SWI 对齐共有乐段，将 A1 排除在训练集之外，同时保留 B，并固定评估文件 A2。

越来越多地受到用户偏好和文本的引导 [2]。连贯且可控的生成需要节奏和长程依赖关系 [3]以及表现力控制，例如情感 [4]。评估对未见作品的泛化能力时，必须考虑因训练期间接触测试集作品而获得的优势。

彼此不重叠的 MIDI 文件可能包含同一作品的不同演奏版本、编曲版本或编辑版本 [5]。因此，测试文件可能未出现在训练集中，但其作品的另一个版本却出现在其中，从而通过暴露降低留出损失（图 1）。Sturm 记录了 GTZAN 中的重复片段和其他版本 [6]；流派研究也表明，艺术家重叠会抬高得分 [7], [8]。

MAESTRO 和 SyMuPe 的 PERiScoPe 会将同一作品的所有演奏放在同一个数据划分中 [9], [10]。 XMusic 的 XMIDI 流程结合文件哈希和色度相似度进行去重 [11], 而 Aria-MIDI 则根据乐曲元数据识别重复作品 [12]。Choi 等人开发了互补的基于内容的 MIDI 去重方法，发布评估代码和筛选后的 LMD 列表，并将跨数据集划分版本识别为一种评估威胁 [13]。 然而，作品

同一性仍难以从不完整的元数据和风格多样的版本中恢复。共同的动机或和声也可能关联不同的作品，因此，隔离必须区分与暴露相关的关系和合理的相似性。

评估还需要了解，版本暴露如何改变不同模型系列的生成损失，以及自动清理在多大程度上近似于未见作品评估。遗漏的关系会使测试作品暴露在外；错误的关系则会移除可接受内容。因此，清理后的损失变化同时反映了暴露和训练池构成，需要联合测量关系准确率、保留内容和模型损失。

我们提出多视图对齐感知同作品识别（MVA-SWI），将互补表征与时间段对应关系相结合，以区分不同版本和偶然相似性。在评估文件固定的情况下，我们测量暴露效应，并将自动隔离的训练集与基于独立作品标签构建的参考集进行比较。

我们的贡献有三点：

- **评估敏感性**。 我们在受控训练条件下量化依赖模型的暴露效应（第 III-B 节）。在 100 对经人工验证的高相似度配对中，用测试作品版本替换 3.29% 的训练窗口，相较于匹配的无关内容，使每组均值负对数似然最多降低 30.16 %（95% 置信区间：27.35–33.02%）。
- **基于内容的隔离**。 MVA-SWI 无需可靠的目标语料库作品标识符即可恢复版本关系。我们还审计了七个 MIDI 合集中检测到的跨数据集划分暴露。
- **下游验证**。 我们根据独立标注的参考数据，评估残余暴露、干净内容保留率以及损失偏差。在 ATEPP 和 LMD-clean 上的两个生成器中，在所评估的非参考条件下，MVA-SWI 的作品级相对负对数似然偏差绝对值均值最小。 一项预算匹配的 ATEPP研究增加了一个独立于 MVA-SWI 选取的参考数据集；另一项独立研究报告了听众评分和自动暴露诊断结果。

II. 相关工作

A. 数据重复与评估可靠性

去重会影响训练语料库构成和评估有效性。 在 GTZAN 上，重复录音对流派分类器的影响并不均衡，而艺术家筛选改变了它们的排名 [14]。 在语言建模中，Lee等人 [15] 使用 n-gram 重叠比较了精确子串匹配与近似文档匹配，将重复与记忆联系起来，并将训练集–验证集重叠与乐观评估联系起来。因此，相关的数据重复粒度取决于预期的泛化目标。

文件哈希和归一化事件指纹能够捕捉精确重复或编码级重复，而演奏版本和编曲版本即使音符有所变化，也可能保留作品的同一性 [13]。Aria-MIDI [12] 通过作曲家与作品标签（包括作品号与曲目编号）处理作品重复问题。这些标签的可用性和一致性限制了基于元数据的整理。我们通过残余暴露、干净令牌保留率，以及与独立标注参考数据之间的损失偏差，研究基于内容的隔离对评估的影响。固定评估内容和匹配随机移除有助于将暴露和选择效应与训练集规模区分开来。

B. 符号音乐检索与去重 *ion*

学习得到的表示使符号音乐的检索不再局限于字面事件匹配。CLaMP [16] 将小节级符号编码与语言–音乐对比学习相结合，使乐谱与描述映射到共享空间。CLaMP 3 [17]将这一方向扩展至符号音乐、音频和多语言文本。这些表示可检索具有共同音乐属性的文件。要识别同一作品的不同版本，还需区分持久的音乐素材与偶然的相似之处。

Choi 等人 [13]提出 CAugBERT，将其预测结果与 CLaMP 相结合，并通过连通分量对重复项进行分组。他们已考虑歌曲的不同版本和跨数据划分泄漏，并发布了评估代码和筛选后的 LMD 列表。在这一作品级研究范围的基础上，MVA-SWI 将有序乐段证据纳入验证，并在固定留出文件的情况下评估数据划分隔离。我们共同衡量残余暴露、合格词元保留率，以及相对于独立标签基准的生成损失偏差，并通过受控训练量化提供测试作品版本的影响。

C. 音乐版本匹配与局部对齐 *ment*

Serrà 等人 [18] 利用色度相似度和局部对齐来检索版本匹配，并通过间隔惩罚适应结构变化。Raffel [5]将基于嵌入的剪枝与动态时间规整相结合，用于音频–MIDI 匹配。ByteCover3 [19]利用局部音频嵌入、近似检索和局部特征重排序，识别短查询，从而减少无关片段的干扰。其 MaxMean 分数对分段最大相似度取均值，但不强制要求时间顺序。

数据集隔离必须控制漏检版本造成的残余暴露，以及错误链接导致的可接受内容移除。MVA-SWI 使用保持顺序的片段路径验证配对、校准判定，并按连通分量对接受的关系进行分组。每条路径内的顺序受到约束；不同区域则可以采用不同的全局顺序。数据划分隔离不会为每个分组保留一个代表项，而是移除与固定验证集和测试集关联的训练文件。残余暴露和不必要的移除用于衡量由此产生的错误。

III. 动机与问题 formulation

A. 从文件互斥到作品独立性

令 $\mathcal{D}_{\mathrm{tr}}$、 $\mathcal{D}_{\mathrm{va}}$ 和 $\mathcal{D}_{\mathrm{te}}$ 分别表示训练集、验证集和测试集 MIDI 文件。我们评估对此前未见过的作品的生成：即演奏版本和版本所基于的作品。令 $g(x)$ 表示文件 x 的参考作品同一性（如有），

表 I：C在固定测试内容下进行受控暴露。U/R/S 表示U无关/R重复-无关/S相同。最后一列为组内-组U至S 降幅(1)，配对95% 工作-组自助法置信区间，以三个训练种子和选定的高-相似度队列为条件。数据集标签定义 R；两位审阅者核验每个提名配对。– 表示不可用条件。比较均在每个模型和目标内进行。

		平均损失				控制对比		相对效应
模型	目标	Clean	无关（U）	重复（R）	Same（S）	R−U	R−S	降幅 [95% CI](%)
Transformer-S	NLL	0.816	0.813	0.817	0.657	+0.004	+0.160	20.30 [17.62, 23.13]
Transformer-M	NLL	0.812	0.800	0.810	0.638	+0.010	+0.172	21.46 [18.65, 24.44]
Transformer-L	NLL	0.798	0.791	0.816	0.589	+0.025	+0.227	27.09 [23.75, 30.58]
MIDI-GPT	NLL	0.736	0.732	0.747	0.526	+0.015	+0.221	30。16 [27.35、33。02]
LSTM	NLL	1.220	1.210	1.237	1.153	+0.027	+0.084	4.74 [3.69, 5.88]
TCN-384	NLL	1.235	1.234	1.237	1.184	+0.003	+0.053	4.04 [3.40, 4.75]
条件变分自编码器	NLL	0.860	0.848	–	0.839	–	–	1。04 [0.44, 1。67]
潜在扩散	MSE	0.015	0.015	–	0.015	–	–	0.01 [−0.07, 0.10]

可用时，$\mathcal{W}(\mathcal{D}) = \{g(x) : x \in \mathcal{D}\}$。文件不相交性要求 $\mathcal{D}_{\mathrm{tr}} \cap \mathcal{D}_{\mathrm{te}} = \varnothing$；作品独立性要求$\mathcal{W}(\mathcal{D}_{\mathrm{tr}}) \cap \mathcal{W}(\mathcal{D}_{\mathrm{te}}) = \varnothing$。如果某个测试作品的任一版本出现在训练集中，即使它们位于不同文件中，该测试作品也属于暴露状态。评估已知作品的新演奏版本时，共享身份也可能是有意为之。参考标签使用语料库作品标识符和曲目元数据，独立于检测器输出和下游损失（标注规则和审计范围：附录 C-C）。

令 $\ell_w^{(c)}$表示条件 c在作品 w的固定测试内容上的损失。负对数似然按每个文件的有效计分词元取均值，再对每部作品的文件等权平均，最后对配对训练随机种子取均值。其他目标保持评分单位和评估内容固定。受控暴露在相同的替换位置和词元预算下，提供另一个测试作品版本（c =同一作品）或无关内容（c = 无关）。对于正的基线损失，平均相对暴露优势为

$$R = \frac{100}{|\mathcal{W}_{\mathrm{te}}|} \sum_{w \in \mathcal{W}_{\mathrm{te}}} \frac{\ell_w^{(\mathrm{unrelated})} - \ell_w^{(\mathrm{same})}}{\ell_w^{(\mathrm{unrelated})}}, \tag{1}$$

其中 $\mathcal{W}_{\mathrm{te}} = \mathcal{W}(\mathcal{D}_{\mathrm{te}})$。 每个受控队列单元都是经确认的同一作品测试集—供体配对；该平均值适用于此选定队列。 相对变化在固定模型和损失定义内取平均。

B. 暴露效应的受控证据

在固定的 LMD 队列中，38,374 个训练窗口里有 1,264 个（3.29%）位于固定替换位置。Clean 保留这些位置的内容，Unrelated 替换为无关内容，Same 则为每个测试作品提供另一个版本。Unrelated 和 Same 的位置及输入词元总数相同。评估固定为 100 组，每组包含一个测试文件和一个供体、三个对应的训练随机种子，以及 20,000 个优化步数。一个冻结的 CAugBERT ∩ CLaMP 图（两个阈值均为 0.99 ）筛选出高相似度配对。两位审阅者确认这 100 组均为同一作品配对，其中包括六组完整作品—节选配对（附录 C-C）。这个选定队列展示了暴露敏感性；它并不能估计所有作品或版本类型的平均效应。 一个基于标签的重复无关对照，在重复次数、长度和音符密度方面进行匹配，以分离重复效应。

暴露效应因模型而异（表 I）。相较于 Unrelated，Transformer-S/-M/-L 的组内负对数似然均值降幅分别为 20.30%、21.46% 和 27.09%，MIDI-GPT 为 30.16%，LSTM/TCN 则为 4.74%/4.04%。 MIDI-GPT 的配对 95% 置信区间为 [27.35, 33.02] %；扩散的均方误差降幅为 0.01% [−0.07, 0。10]。变分自编码器解码器负对数似然和扩散噪声均方误差探测的是不同目标，因此这些数值不能用于对模型系列的总体暴露敏感性进行排名。MIDI-GPT 在 5,000 和 10,000 次更新下的额外运行分别得到 4.39% 和 14.44% 的降幅（附录 C-C ）。

全部六种自回归配置在 Same 条件下的损失都低于任一无关对照。例如，MIDI-GPT 在重复无关条件下的负对数似然为 0.747，而在 Same 条件下为 0.526；在匹配重复程度后，仍保留了暴露优势。

C. 关系恢复与隔离标准

当 g 不可用时，我们根据 MIDI 内容估计同一作品关系$r(x_i, x_j) = \mathbf{1}[g(x_i) = g(x_j)]$ 。据此定义分组 $\hat{g}$，并在 g 可用时与之对照评估；对于文件集 S，令 $\hat{g}(S) = \{\hat{g}(x) : x \in S\}$。我们通过保留内容和后续模型损失来评估隔离效果。

对于保留的训练集 $\mathcal{D}_{\mathrm{tr}}^{\mathrm{iso}}$，独立标签用于衡量残余暴露：

$$E_{\mathrm{work}} = \frac{|\mathcal{W}(\mathcal{D}_{\mathrm{tr}}^{\mathrm{iso}}) \cap \mathcal{W}_{\mathrm{te}}|}{|\mathcal{W}_{\mathrm{te}}|}. \tag{2}$$

为衡量不必要的移除，定义可纳入的训练子集 $\mathcal{D}_{\mathrm{safe}} = \{x \in \mathcal{D}_{\mathrm{tr}} : g(x) \notin \mathcal{W}(\mathcal{D}_{\mathrm{va}} \cup \mathcal{D}_{\mathrm{te}})\}$。若$t(x)$表示在固定分词器和预处理方案下可用的训练词元数量，且 $\sum_{x \in \mathcal{D}_{\mathrm{safe}}} t(x) > 0$，则干净令牌保留率为

$$K_{\mathrm{token}} = \frac{\sum_{x \in \mathcal{D}_{\mathrm{tr}}^{\mathrm{iso}} \cap \mathcal{D}_{\mathrm{safe}}} t(x)}{\sum_{x \in \mathcal{D}_{\mathrm{safe}}} t(x)}. \tag{3}$$

遗漏的关系可能留下暴露；错误关系可能降低干净令牌保留率。

对于下游评估，在独立的标签隔离数据上训练参考模型。对于每个参考损失为正的测试作品，定义

$$\delta_w = \frac{\ell_w^{(\mathrm{iso})} - \ell_w^{(\mathrm{ref})}}{\ell_w^{(\mathrm{ref})}}. \tag{4}$$

对于非空测试作品集，带符号偏差 B_{signed} 是 $100\delta_w$ 的作品级均值，绝对偏差为

$$B_{\text{abs}} = \frac{100}{|\mathcal{W}_{\text{te}}|} \sum_{w \in \mathcal{W}_{\text{te}}} |\delta_w|. \tag{5}$$

绝对偏差会阻止不同作品之间相互抵消。结果一致性取决于训练集组成和数据量；匹配随机移除有助于区分关系引导的选择与数据缩减。

IV. 用于作品级隔离的多视图对齐

MVA-SWI 通过恢复版本关系，将训练内容与固定的验证集和测试集隔离开来。即使其他方面发生变化，低音模式、和声或旋律动机仍可能保持不变。我们通过互补视图检索候选项，验证有序的局部对应关系，并对校准后的成对决策进行分组（图 2）。推理使用全部三个分区中的 MIDI 内容；作品标签仅用于开发和评估。算法 1 概述了冻结验证器推理；附录 A 详细说明了解析和特征回退方案。

A. 多视图候选检索

每个文件由五种视图表示：低音音程 n 元组、移调规范化和声序列、上声部轮廓动机，以及音程和时值直方图。令牌视图通过非负哈希、次线性词频和逆文档频率（IDF）编码重复出现的模式；直方图则捕捉音高和时序分布。所有视图均经过 ℓ_2 归一化。IDF 会根据当前检索语料库进行调整，无需作品标签，而验证器保持固定（解析、量化和加权：附录 A）。

Faiss [20]会为每个视图构建分层可导航小世界（HNSW）索引 [21]，并查询每个有效文件。Token 向量会经过投影和重新归一化；直方图则直接建立索引。内积搜索可近似实现余弦近邻检索。未投影的向量用于提供全局验证特征；附录 A 说明了投影和索引参数。

算法 1MVA-SWI 核心推理与作品隔离**要求：** 固定分区（$\mathcal{D}_{\text{tr}}$，$\mathcal{D}_{\text{va}}$，$\mathcal{D}_{\text{te}}$）**要求：** 冻结的 f_1，f_2，f_3，校准参数（a，b_0），阈值 τ **确保：** 推断组 $\hat{g}$ 和保留的训练集 $\mathcal{D}_{\text{tr}}^{\text{iso}}$1：$\mathcal{D} \leftarrow \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{va}} \cup \mathcal{D}_{\text{te}}$2：构建五个检索视图；在 $\mathcal{D}$ 上拟合逆文档频率（附录 A） 3：为每个视图构建一个 HNSW 索引；为全局有效文件检索前 100 个非自身近邻 4：依据（6），从任一单一视图中的互为前 50 的配对构成 $\mathcal{C}$；每对仅存储一次，记为 $i < j$5：**对于** 每个 $\{i, j\} \in \mathcal{C}$ **执行**6：计算 29 个全局特征和 8 个检索特征 7：通过零移位对齐（附录 B-A）计算 18 个路径特征和两个辅助色度特征 8：按照附录 A-B 中的顺序组装 $z_{ij} \in \mathbb{R}^{57}$ 9：$\hat{}p_{ij} \leftarrow \sigma\big(a\,\frac{1}{3}\sum_{b=1}^{3} f_b(z_{ij}) + b_0\big)$10：**结束循环** 11 $\mathcal{E}^+ \leftarrow \{\{i, j\} \in \mathcal{C} \;\; p_{ij} \geq \tau\}$：：^：查找（$\mathcal{D}$，$\mathcal{E}^+$）的连通 12 分量，并依据（10）隔离训练文件 13：**返回** $\hat{g}$，$\mathcal{D}_{\text{tr}}^{\text{iso}}$

令 $N_v^k(i)$ 表示 HNSW 为视图 v中的文件 x_i 返回的前 k 个非自身邻居。仅当两个文件在相同视图中相互检索到对方时，我们才保留该文件对：

$$\mathcal{C} = \bigcup_{v \in \mathcal{V}} \big\{\{i, j\} : i < j,\; j \in N_v^{50}(i),\; i \in N_v^{50}(j)\big\}. \tag{6}$$

两种检索方向必须在同一个视图中出现；任一视图均可。每个查询最多保留 100 个非自身邻居：前 50 个用于确定候选项，全部邻居均用于提供验证等级。各视图的互选集合合并后去重，并按固定的输入文件索引为每对结果仅存储一次。训练、校准和推理均遵循此规则。

B. 保序局部对齐

为检验共享素材是否构成连贯的音乐片段，我们比较一系列短分段。将第一个音符的起始时间移至零后，我们用文件头中的每四分音符脉冲数（PPQ）除 MIDI 时钟刻度，并以两个 MIDI 四分音符单位（2×PPQ时钟刻度）为单位划分分段。此方法既不使用速度图，也不推断演奏版本的拍。最多保留前 512 个分段。每个分段记录高音和低音音高在八个起始时间箱中的分布、起音计数向量、音级轮廓以及音符数。每个音符仅计入其起始时间所在的分段，其音级权重为力度乘以时长（上限为两个四分音符单位）。

对于分段 u和 v，旋律相似度 m_{uv} 是对应分箱中包含高音区 MIDI 音高 p 和 q的分箱的 $\exp(-|p-q|/2.5)$ 均值。低音相似度 b_{uv} 以类似方式计算；若没有共同有效的分箱，则任一相似度均为零。和声相似度 h_{uv} 和节奏相似度 r_{uv} 分别是经过 ℓ_2归一化的音级向量与起始计数向量的内积。给定音符数 a_u 和 a_v，密度相似度为 $d_{uv} = \exp(-|\log(1+a_u) - \log(1+a_v)|)$。我们将它们组合为

$$s_{uv} = 0.30m_{uv} + 0.15b_{uv} + 0.25h_{uv} + 0.20r_{uv} + 0.10d_{uv}. \tag{7}$$

局部对齐采用观测到的音高，不进行音高移位。零移位指的是音高，而非速度。对齐允许段落移位和间隙，但不进行显式的速度归一化；弹性速度和速度差异仍可能损害对应关系（附录 C-B）。全局移调不变表示和汇总色度证据提供了独立的验证特征。

我们通过仿射间隙动态规划 [22], [23]恢复局部对应关系。令 $q_{uv} = 2(s_{uv} - 0.525)$，令 H_{uv}为以 (u, v) 结束的最佳局部得分，并令 E_{uv} 和 F_{uv}为两个间隙状态。从零边界状态开始，

$$\begin{aligned} E_{uv} &= \max\{0, H_{u-1,v} - \gamma_o, E_{u-1,v} - \gamma_e\}, \\ F_{uv} &= \max\{0, H_{u,v-1} - \gamma_o, F_{u,v-1} - \gamma_e\}, \\ H_{uv} &= \max\{0, H_{u-1,v-1} + q_{uv}, E_{uv}, F_{uv}\}, \end{aligned} \tag{8}$$

其中 $\gamma_o = 0.35$ 和 $\gamma_e = 0.08$。从得分最高的状态开始回溯，可得到由匹配分段对组成的路径。允许从零开始意味着匹配可以从任一文件内部开始，而间隔状态则用于容纳中间未匹配的分段。

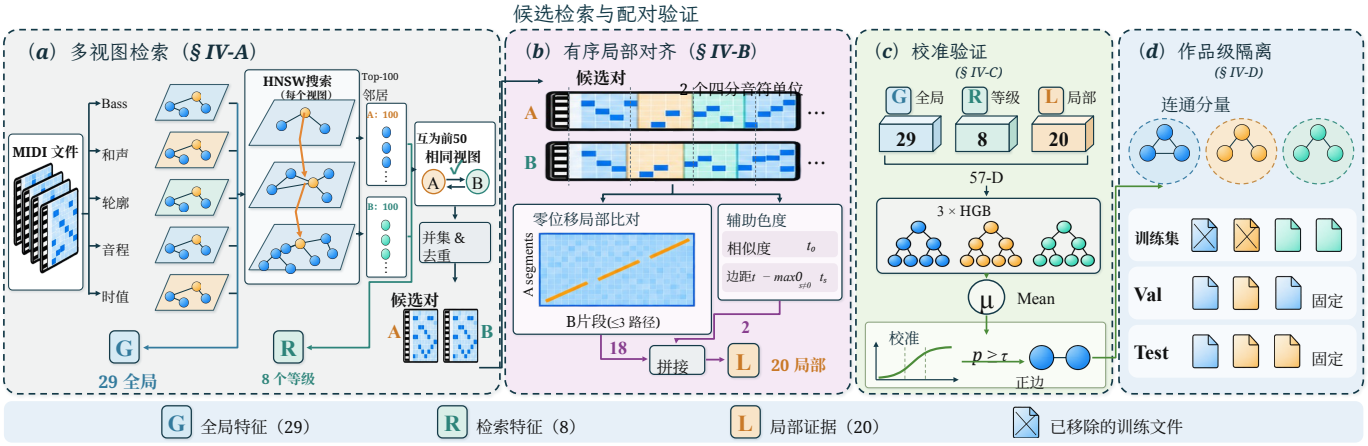


图 2。MVA-SWI 工作流程概览。(a) HNSW 针对每种视图最多检索 100 个邻居，以生成排名特征；在相同视图中互为前50名邻居的配对会跨五种视图合并。(b) 零位移局部比对提供 18 个路径特征，并辅以两个色度辅助特征。(c) 分类器集成对 57 维证据向量进行评分，并通过 Sigmoid 校准进行校准。(d) 阈值为正的边所构成的连通分量用于指导训练数据移除，同时验证集和测试集保持固定。

我们最多提取三条路径。如果得分最高的路径包含少于四个匹配对，则终止提取。否则，在下一次搜索前会将该路径匹配的行和列屏蔽。每条路径都保留时间顺序，而不同路径可能处于不同的全局顺序中。对于保留序列长度 n_A，n_B 和匹配索引集合 I_A，I_B，覆盖率值分别为 $c_A = |I_A|/n_A$ 和 $c_B = |I_B|/n_B$。它们的调和均值为 $c_{\mathrm{hm}} = 2c_Ac_B/(c_A+c_B)$；当两个覆盖率均为零时，调和均值为零。覆盖率基于保留的前缀进行测量。附录 B-A 给出了提取过程，包括回溯并列和停止规则。

C. 校准的同作品验证

每个候选项对应一个 57 维向量 z_{ij}：29 项全局比较、8 个检索统计量和 20 个局部特征。全局特征比较音乐模式、分布和统计量；检索特征概括排名和相互支持。局部区块包含 18 个路径导出坐标，以及两个与对齐无关的色度特征：零移位聚合相似度及其相对于最佳非零移位的差值。附录 A-B 定义了所有坐标和边界情况。

目标数据包含来自同一录音的文件，但开发集对每个录音标识符仅使用一个文件，不含重复校验和，并且训练集、校准集和测试集中的作品彼此不重叠。正样本对在不同录音标识符之间共享同一作品；负样本对则来自不同作品。成对基准测试还排除了同一录音配对，因此衡量的是跨录音识别。未标注语料库推理不采用基于录音的排除规则（队列和范围：附录 B-B）。

我们保留所有正候选对，并分别为三个类别平衡的基于直方图的梯度提升分类器独立抽取负样本。令 $f_b(z_{ij}) \in [0,1]$ 为分类器b预测的正类概率。将它们的算术均值直接通过 Sigmoid 函数 [24]进行校准：

$$\bar{f}_{ij} = \frac{1}{3}\sum_{b=1}^{3} f_b(z_{ij}), \qquad \hat{p}_{ij} = \sigma(a\bar{f}_{ij} + b_0), \tag{9}$$

其中 $\sigma(t) = (1+e^{-t})^{-1}$。完整的校准候选池既用于拟合（$a$，$b_0$），也用于选择阈值$\tau$。阈值选择以校准集 F1 值最大化为目标，同时要求精确率至少为 0。97，且预测为正的候选对至少有 20 对。针对指定的特征表示和 HNSW 候选规则拟合分类器、校准器和阈值，然后将其冻结，用于在各目标语料库上进行评估。端到端评估会将检索遗漏的符合条件的同一作品候选对计为假阴性。抽样、优化和阈值并列时的处理规则详见附录 B-B。

D. 关系引导的作品隔离

在所有文件上构建的无向图包含阈值正边 $\mathcal{E}^+ = \{\{i,j\} \in \mathcal{C} : \hat{p}_{ij} \geq \tau\}$。连通分量定义作品组：沿每条正边合并单例组，且与处理顺序无关（附录 B-C）。

令 $\hat{g}$ 表示连通分量分配。在固定验证集和测试文件的情况下，我们保留

$$\mathcal{D}_{\mathrm{tr}}^{\mathrm{iso}} = \{x \in \mathcal{D}_{\mathrm{tr}} : \hat{g}(x) \notin \hat{g}(\mathcal{D}_{\mathrm{va}} \cup \mathcal{D}_{\mathrm{te}})\}. \tag{10}$$

每个接受的关系的两个端点都位于同一推断组中，因此没有阈值正边连接保留的训练集与留出文件。假阳性桥接边可能合并无关作品，而缺失边则可能将同一作品分割开来。残余工作暴露和干净内容保留率使用独立标签，并依据公式 (2) 和 (3) 衡量这些错误。对于未标注语料库，报告 $\hat{g}$ 下的重叠，作为检测到的组暴露。因全局解析而被排除的文件仍属于未解决的单例组。

对于 N 个文件和 V 个视图，相互 Top-k 检索在跨视图去重前最多生成 $VkN/2$ 个无序候选对。HNSW 避免了全量两两相似度扫描；其搜索成本和召回率取决于索引和搜索预算。每轮对齐的时间和内存成本为 $O(n_An_B)$，最多进行三轮。连通分量可通过并查集在 $O((N+|\mathcal{E}^+|)\alpha(N))$ 时间内求得，其中 α 为反阿克曼函数。附录 B-C

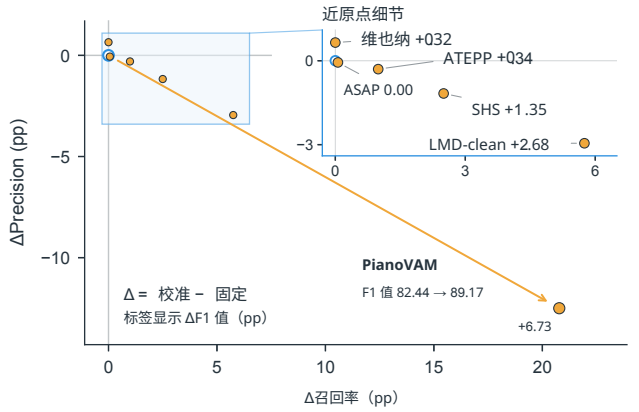


图 3. 校准带来的提升与精确率–召回率权衡。坐标表示校准值减去固定值后的召回率和精确率差（百分点，pp）；数字标签表示相应的 F1 值变化。维也纳表示维也纳 4×22。表V给出了校准后的 95% F1置信区间。

区分了核心检测器与用于七数据集审计的独立配置。

V. 实验

我们在七个 MIDI 合集中评估关系恢复、消融实验和检测到的重叠。由独立标注的研究衡量残余暴露、干净令牌保留率，以及与标签隔离参考的负对数似然一致性；聆听评估用于评估感知到的音乐性，而自动诊断用于衡量与留出内容的相似度。

A. 实验设置

固定阈值比较使用 ATEPP [25],ASAP [26],LMD-clean [13], 和 MAESTRO [9]。校准还涵盖SecondHandSongs（SHS） [27], 维也纳 4×22 [28]和 PianoVAM [29], ，但不包括 MAESTRO。每项比较均固定其队列。我们使用冻结版 v1.9（第 IV 节），在作品互斥的 LMD 分区上拟合：五视图 HNSW检索、同视图互选Top-50候选项、零移位对齐，以及三个包含 57 个特征的分类器。阈值为0。597596；正边连通分量定义隔离组。附录 B-B 详述拟合与配置；附录 B-C 介绍独立语料审计器。

基线方法包括文件SHA-256、归一化事件指纹、CAugBERT、CLaMP，以及 CAugBERT ∪ CLaMP [13], [16], 在 0.93/0.87 的阈值下，只要任一模型预测为正例即予以接受。隔离实验还测试了二者在 0.99 阈值下的交集，校准并集，以及适用情况下的元数据匹配。成对精确率（P）、召回率（R）和 F1 值不计入同一录音配对，并将未检索到的正例计为假阴性。特征和检索结果是对 LMD-clean、ATEPP 和 ASAP 的未加权宏平均。隔离实验使用公式 (2)–(5)；附录 C 定义了区间，以及不同数据划分和分类器子集间的变异。

B. 同一作品识别与校准

暴露控制要求恢复跨版本关系，同时限制可能移除可接受内容。在固定阈值下，MVA-SWI 在 ATEPP、ASAP、LMD-clean 和 MAESTRO 上的 F1 值分别为 97.90%、99.73%、79.67% 和 100.00%（表 II）。除了 LMD-clean 的一部分配对外，精确匹配几乎无法恢复任何跨录音关系。嵌入相似度能够恢复更多正样本，但精确率与召回率的平衡因数据集而异。在 ATEPP 上，CAugBERT ∪CLaMP 的召回率高于任一组成方法，但精确率仅为 22.86%；MVA-SWI 的精确率达到 98.08%，召回率为 97.72%。在 LMD-clean 上，尽管精确率达到 96.18%，MVA-SWI 仍漏掉了 2,037 个参考正例配对中的 652 个，因此有必要在隔离实验中直接测量残余暴露。

我们比较了独立的目标数据集校准与直接迁移固定的 LMD 阈值（图 3）。在全部六个队列中，MVA-SWI 的 F1 点估计均高于经过校准的 CAugBERT ∪CLaMP（表V）。校准使其在 LMD-clean 上的 F1 值从 79.67% 提高至 82.35%，在 PianoVAM 上从 82.44% 提高至 89.17%，而 ASAP 仍为 99.73%。在 PianoVAM 上，召回率从 70.13% 升至 90.91%，精确率则从 100.00% 降至 87.50%。SHS 校准后的 F1 值仅为 72.99%（95% 置信区间：51.28–85.71%），与 MAESTRO 在固定阈值下达到的 100% F1 值形成对比。结合 LMD-clean 上漏检的正例，这限制了我们对模型在异质来源中具有普遍稳健性的结论。

C. 检索、对齐与验证的贡献

候选召回率从仅使用一个所选视图时的 86.16% 提升至使用五个视图时的 96.13%；其中 LMD-clean 的提升最大，从 59.55% 提升至 88.61%（图 4a）。这支持在异构数据中使用互补视图。与部署的 Top-50 预算相比，Top-100 仅使宏平均召回率增加 1.32 个百分点（图 4b）。

在候选项固定的情况下，移除 20 个局部特征会导致特征块 F1 最大幅度下降，从 92.43% 降至 88.32%，并使召回率从 88.45% 降至 82.07%（表 III）。移除全局特征或检索特征会使 F1 分别下降 1.92 或 0.94 个百分点。用互近邻规则替代学习式验证后，精确率为 43.29%，F1 值为 57.30%。 局部区块包含两个与对齐无关的色度特征，因此该消融实验无法单独考察时间顺序。移位与前缀测试使用完整向量（扩展结果）。

将集成中的分类器数量从 3 个增加到 5 个后，对每种规模下的所有组合取均值，宏平均 F1 从 92.59% 变为 92.60%（图 4c）。92.43% 的完整模型结果（表 III）和隔离实验使用核心系统的 3 个固定部署随机种子。 在 512 个分段时，Top-3 移位变体在 LMD-clean 上的召回率提高了 3.63 个百分点，精确率则降低了 2.17 个百分点。在 ATEPP 和 ASAP 上，F1 值分别下降了 0.18 和 0.15 个百分点。将零移位前缀扩展至 768 个分段后，F1 值的变化最多为 0.02 个百分点。对于标注的中位数节拍间隔相差超过 20% 的样本对，ASAP 的召回率为 98.32%；在采用音符时长跨度代理指标时，ATEPP 的召回率为 94.50%。附录 C-B 报告了分层结果和固定验证器参数敏感性。

*D. MIDI*合集中的检测到的暴露

我们审查了七个冻结语料库快照：POP909 [30],ASAP、PDMX [31],GigaMIDI [32],MAESTRO、LMD [5],

表 II 固定工作点下的成对识别。P、召回率和 F1 值均以百分比表示；FN 统计漏掉的参考正例配对，包括未检索到的样本对。CAugBERT 和 CLaMP 分别使用 0.93 和 0.87 的阈值；∪ 表示二者的预测并集。MVA-SWI 使用 0.597596。当没有预测出正样本对时，精确率未定义，以破折号表示。

	ATEPP			ASAP			LMD-clean			MAESTRO		
方法	P	R	F1 值 FN	P	R	F1 值 FN	P	R	F1 值 FN	P	R	F1 值 FN
SHA-256	–	0.00	0.00 3339	100.00	0.02	0.04 4987	–	0.00	0.00 2037	–	0.00	0.00 880
事件指纹	–	0.00	0.00 3339	100.00	0.02	0.04 4987	96.72	20.27	33.52 1624	–	0.00	0.00 880
CAugBERT	12.69	31.84	18.14 2276	35.67	50.36	41.76 2476	94.09	28.92	44.24 1448	62.45	38.18	47.39 544
CLaMP	62.71	69.54	65.95 1017	86.64	30.03	44.60 3490	94.64	26.02	40.82 1507	88.45	53.98	67.04 405
CAugBERT ∪CLaMP	22.86	76.61	35.21 781	40.34	64.43	49.61 1774	94.13	33.04	48.91 1364	70.48	70.00	70.24 264
MVA-SWI	98.08		76	99.84		19			652			0

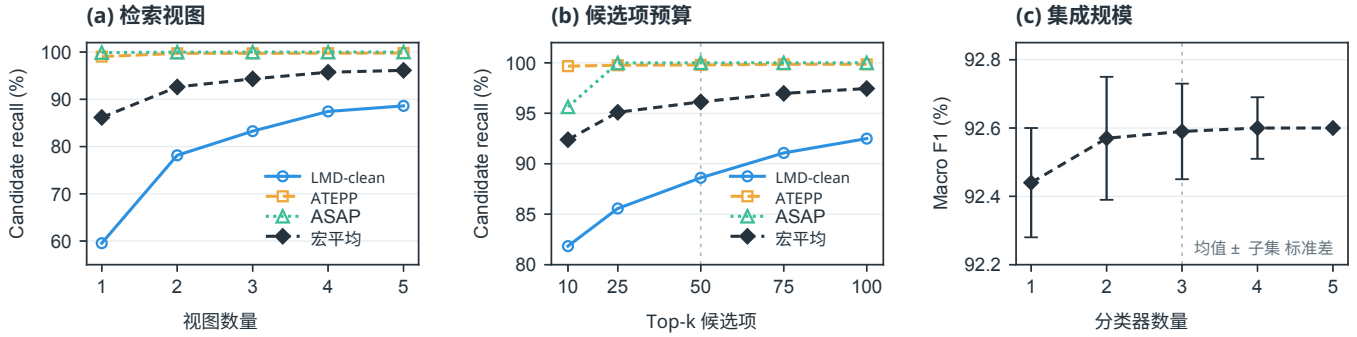


图 4. 三个数据集按数据集划分的检索覆盖率和宏平均集成 F1 值。(a) 对于每种规模，视图子集均在 MAESTRO 验证集上独立选取。(b) 随着邻居预算 Top-k 的增加，候选召回率的变化。(c) 误差线表示五个训练后的分类器的所有子集上的总体标准差，并且针对每个子集重新拟合校准。在五个分类器处的零离散度反映了只有一种组合。

表 III 验证器消融实验：LMD-clean、ATEPP 和 ASAP 上的宏平均百分比。 ✓/×: 使用/省略；CLF.：分类器；D：输入维度。 †：在至少两个低音/和声/动机视图中获得互为前 100 的支持；候选项仍为同视图互为前 50。特征移除变体经过重新训练和重新校准。平均精度使用候选项；召回率包括未检索到的正例。

验证器组件				宏平均演奏版本（%）				
全局	检索	局部	Clf.	D	P	R	F1	AP
✓	✓	✓	✓	57				
×	✓	✓	✓	28	98.37	85.14	90.51	94.99
✓	×	✓	✓	49	97.68	87.37	91.49	95.27
✓	✓	×	✓	37	97.99	82.07	88.32	94.53
×	✓†	×	×	–	43.29	88.45	57.30	75.59

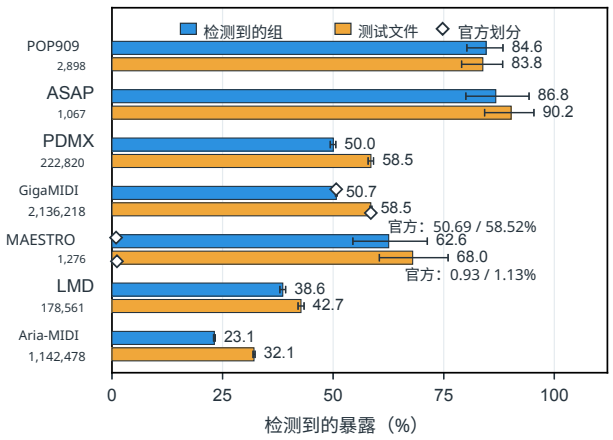


图5. 七个 MIDI 合集中的检测到的暴露。柱状图显示 200 次随机 80/10/10 文件划分下组/文件暴露度的均值；误差线表示第 2.5–97.5 经验百分位数。菱形标记和组/文件标签表示官方划分；名称下方的计数表示输入文件数量。音程量化了在推断组条件下划分的变异性。

和 Aria-MIDI [12], , 共计 3,685,318 个输入文件。近似检索、初步门控以及独立训练的 49特征验证器在各语料库中保持固定（附录 B-C）。此配置不同于 57特征基准检测器，用于衡量预测分组之间的重叠。

如果测试组中包含训练文件，则该组存在暴露；测试文件暴露度根据各预测分组中的测试文件数量进行加权。在随机文件划分下，检测到的组暴露度均值从 Aria-MIDI 的 23.13% 到 ASAP 的 86.77% 不等，测试文件暴露度均值则从 32.11% 到 90.23% 不等（图5）。GigaMIDI 的官方划分产生了 50.69% 的组暴露度和 58.52% 的文件暴露度，接近随机划分下 50.71% 和 58.52% 的均值。MAESTRO 的官方划分仅产生 0.93%/1.13%，而随机划分下则为 62.56%/67.98%。

因此，文件分区中仍可能存在大量检测到的重叠，而划分构造可以减少这种重叠。 在包含三个队列的已标注基准测试中，49-featureFast配置的配对精确率为 91.70–99.70%，召回率为 62.89–99.66%；与 v1.9 相比，F1 值差距最大的是 LMD-clean（74.61% 对 79.67%）。在对四个未标注语料库进行的部署审计中，审阅者确认每个语料库随机抽样的 100 条接受的边中有 91–95 条属实，并确认每个语料库定向抽取的 20 条目标桥接边中有 15–17 条属实。在选出的 20 个大型连通分量中，有 9 个包含经确认来自多个作品的成员。附录 C-D 报告了这些检查、

表 IV默认和校准基线方法在配对随机划分上的平均结果。 E 为残余工作暴露(%，越低越好）； K 为 Clean-令牌保留率(%，越高越好）。训练集 ∪ 和 ∩ 基线方法分别使用并集和交集，。； 阈值对依次列出CAUGBERT 在前，CLAMP 在后。– 表示未报告的元数据基线。

	MVA-SWI (0.597596)		CAugBERT ∪ CLaMP 默认 (0.93/0.87)		CAugBERT ∩ CLaMP 交集 (0.99/0.99)		CAugBERT ∪ CLaMP 校准 (0.964/0.876)		元数据匹配	
数据集	$E\downarrow$	$K\uparrow$	$E\downarrow$	$K\uparrow$	$E\downarrow$	$K\uparrow$	$E\downarrow$	$K\uparrow$	$E\downarrow$	$K\uparrow$
ATEPP	1.24	99.59	18.60	31.35	99.47	100.00	44.74	75.57	22.56	95.56
ASAP	1.61	98.82	34.81	42.66	90.69	100.00	72.54	89.18	–	–
LMD-clean	17.75	99.19	35.66	99.35	49.74	99.90	36.49	99.43	–	–
MAESTRO	0.00	100.00	20.46	66.56	67.63	100.00	41.94	84.51	2.01	100.00

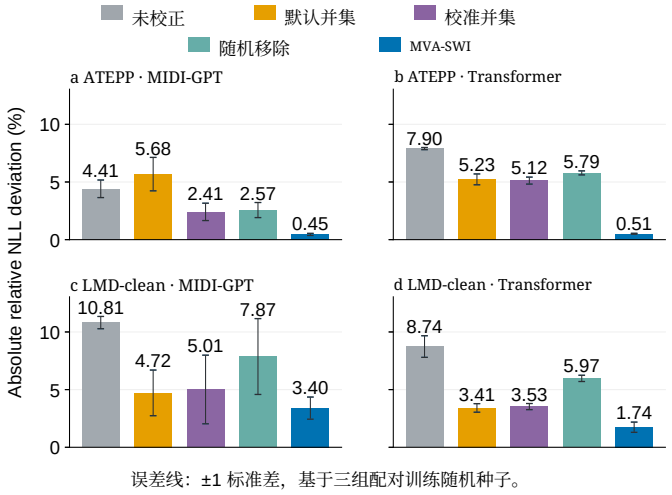


图 6。 相对于每首作品的标签参考的绝对相对负对数似然偏差 B_{abs}（越低表示越接近），包括校准并集基线。柱形表示均值，误差线表示 ±1 三个配对训练随机种子下的标准差。训练集池详情以及与图 7 的区别见附录 C-E。

解析失败情况以及端到端审计成本。

E. 隔离质量与下游评估

我们通过独立标注的暴露与保留率，以及与标签隔离参考项的损失一致性来评估移除效果。

残余暴露与内容保留率。在配对的随机 80/10/10 划分下，仅清理训练池；针对三个主要来源各进行 20 次划分（附录 C-C）。在固定工作点，MVA-SWI 在全部四个数据集中保留至少 98.82% 的干净词元；在 ATEPP、ASAP、LMD-clean 和 MAESTRO 上的残余暴露率分别为 1.24%、1.61%、17.75% 和 0.00%（表 IV）。在每个数据集中，校准并集的干净令牌保留率都高于默认并集，但残余暴露更多：ATEPP 上达到 44.74%，ASAP 上达到 72.54%。元数据匹配在 ATEPP/MAESTRO 上分别留下 22.56%/2.01% 的暴露。高阈值交集保留了更多内容，但漏掉了更多测试作品版本。MVA-SWI 在所有情况下的暴露率最低，但其 LMD-clean 保留率略低于嵌入基线。

留一来源研究测试阈值迁移。开发集暴露目标为 1% 时，在留出的 LMD-clean 上，暴露率为 12.95%，干净令牌保留率为 92.63%

来源；有些更严格的目标不存在可行阈值（附录 C-D）。因此，迁移阈值时必须评估目标队列暴露率和保留率。

与标签隔离参考的一致性。使用不同于上述重复划分平均值的固定队列，我们在 ATEPP 和 LMD-clean 上训练 MIDI-GPT 和 Transformer，采用六种条件：未校正、匹配随机移除、MVA-SWI 隔离、默认和校准 CAugBERT ∪ CLaMP 隔离，以及独立标注参考。每种条件均使用三个训练随机种子，并在每个数据集中固定评估内容。随机移除与 MVA-SWI 匹配，分别保留 ATEPP 中的 486 个文件和 LMD-clean 中的 1,023 个文件，从而将内容选择与文件数量区分开来。其他条件保留的数据量各不相同（附录 C-E）。一致性通过式 (5) 中每首作品相对 NLL 偏差的绝对值B_{abs}来衡量；数值越小表示一致性越高。

在所有四种数据集—模型设置中，MVA-SWI 的均值 B_{abs} 在五种非参考条件中最低（图 6）。在 ATEPP 上，MIDI-GPT 和 Transformer 的偏差分别为 0.45% 和 0.51%，而匹配随机移除条件下分别为 2.57% 和 5.79%。MVA-SWI 未暴露 121 个测试作品中的任何一个，并保留了 98.91% 的干净词元；随机移除则暴露了 72 个作品，并保留了 63.97%。在 LMD-clean 上，MVA-SWI 对应的偏差分别为 3.40% 和 1.74%，而随机移除条件下分别为 7.87% 和 5.97%。它保留了 99.67% 的干净词元，同时暴露了 142 个测试作品中的 25 个。校准并集在 ATEPP 上的结果为 2.41%/5.12%，在 LMD-clean 上为 5.01%/3.53%（依次对应 MIDI-GPT/Transformer）。这些均值比较包含随机种子标准差，但不能据此确认不同条件之间存在显著差异。

控制训练预算。一项独立的六条件 ATEPP研究对训练更新步数、窗口和计分词元进行了匹配（图 7 和附录 C-E）。在 ATEPP 标签下，MVA-SWI 未暴露 121 件测试作品中的任何一件，其相对负对数似然偏差均值为 $+2.15\%$ ，相对于独立选定的参考值（配对 95% 置信区间 $[+1.85\%\ +2.48\%]$）。图 6 和图 7 使用不同的训练池、参考值和预算方案。它们的绝对偏差和带符号偏差属于不同的估计目标，而不是对同一批运行结果的替代性汇总。

F. 感知质量与生成诊断

我们使用 MIDI-GPT 和 Transformer，在 ATEPP 和 LMD- clean 上描述五种下游条件的听众评分，并单独报告自动续写诊断结果（图 8 和表X）。这五种条件不包括

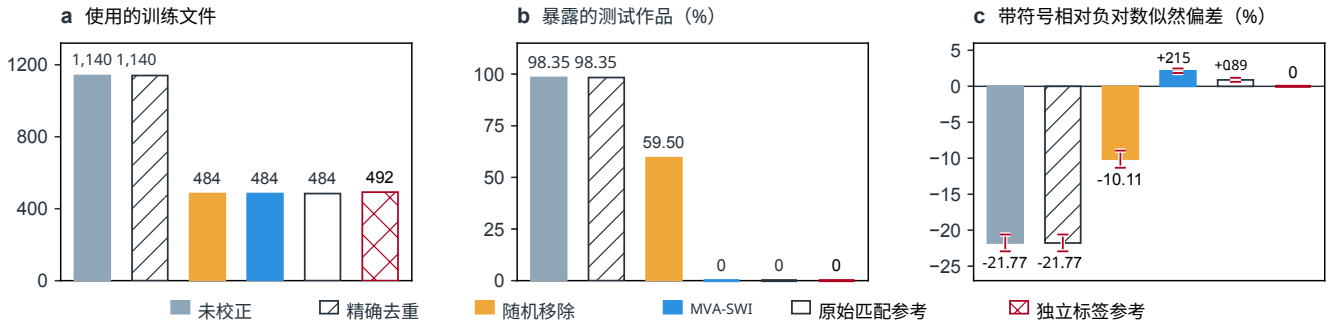


图 7. 受预算控制的 ATEPP 隔离：(a) 训练文件；(b) 在 ATEPP 标签下暴露的测试作品（%）；以及 (c) 相对于独立标注参考的每首作品相对 NLL 偏差均值（%），并附配对 95% 置信区间。负偏差表示损失更低。每个条件均使用 20,000 次更新、11,316 个训练窗口和 5,884,369 个计分词元。

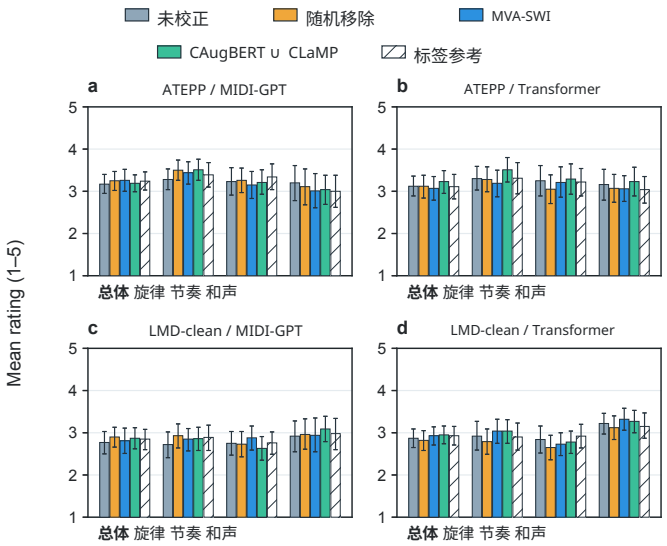


图 8. 盲评评分（1–5；越高越好）。柱状图显示总体音乐性（主要指标）、旋律连贯性、节奏自然度和和声连贯性的均值及 95%置信区间。

新加入的校准并集。本研究采用十个训练随机种子，与隔离实验所用的三个不同；聆听评估提示则取自与清理实验相同的测试集。附录 D 详述了刺激材料分配和统计分析。

盲听。80 位听众参与评估，其中一半有音乐背景，另一半没有，共完成 4,800 次正式音频评估。听众只听了生成的续写；这些续写由 MIDI 经 FluidSynth 渲染，速度为每分钟 120 拍，采样率为 16 千赫。他们在为片段评分前均已提供知情同意。每段匿名的八秒片段均按五分制评定整体音乐性、旋律连贯性、节奏自然度和和声连贯性。图 8 报告了各条件的均值和 95% 置信区间；配对对比以整体音乐性为主要结局。ATEPP 上 MVA-SWI 与未校正条件的差异为 $+0.09$ ($[-0.09, +0.27]$)，分别对应 MIDI-GPT 和 -0.05($[-0.25, +0.12]$)，对应 Transformer。在 LMD-clean 上，差异分别为$+0.05$ ($[-0.14, +0.23]$) 和 $+0.06$ ($[-0.12, +0.23]$)。这些 95% 置信区间是在给定十个训练种子的条件下，对听众和作品差异进行估计所得。对比结果为

描述性地表达，并保留一系列与数据相符的正面和负面影响。

自动诊断。自动评估采用固定的 120 部 ATEPP 作品和 119 个 LMD-clean 参考组。旋律相似度和节奏相似度衡量与真实测试集续篇的一致性；FMD-CLaMP1 将弗雷歇音乐距离 [33]应用于生成的和共享的自然续篇的 CLaMP嵌入 [16]。在四种数据集—模型设置中，MVA-SWI 将长段落匹配率降至 2.46–4.06%，而未校正组为 21.90–23.89%，匹配随机移除组为 10.48–11.69%；标签隔离率为 2.41–3.94%（表X）。其旋律和节奏相似度均值分别与参考值相差不到 0.44 和 0.34 个百分点，FMD 相差不到 0.18（扩展结果）。在每种设置中，未校正组的续篇相似度更高、FMD 更低。由于比较以留出续篇为目标，这些测量结果反映的是与该目标的一致性，也可能受到作品暴露的影响。听众评分则提供了独立的感知结果。

VI. 结论

MVA-SWI 旨在解决符号音乐评估中的一个关键挑战：同一作品的不同演奏版本或编曲可能同时出现在训练集和测试集中。它结合互补的音乐表示、时间对齐和校准验证，识别这些版本，并在保持评估文件固定的同时将其从训练集中移除。

我们对 100 对经人工验证且相似度较高的样本进行的受控研究表明，这一点为何重要：接触同一作品的另一版本会降低预测损失，且影响因模型而异。在 ATEPP 上匹配训练预算后，MVA-SWI 移除了数据集标签所识别出的所有暴露，平均相对测试损失比独立选取的参照高 2.15%。但在 LMD-clean 上，隔离仍不彻底，部分暴露依然存在。在所有四种评估设置中，生成的续写所包含的长段落匹配也更少。然而，聆听结果仍不足以确定音乐质量是否存在差异，也无法证明二者等效。

漏检版本可能导致重叠未被发现，而错误的链接则可能移除有效的训练材料。语料库范围估计还取决于审计者的覆盖范围，并且检测到的

重叠本身并不能证明侵犯版权。这些限制进一步表明，有必要在报告预测损失的同时，报告剩余重叠、保留的训练内容以及音乐质量。总之，我们的研究结果支持在评估对未见作品的泛化能力时，将每部作品的各个版本保留在同一数据划分中。

参考文献

[1] C.-Z. A. Huang、A. Vaswani、J. Uszkoreit、N. Shazeer、I. Simon、C. Hawthorne、A. M. Dai、M. D. Hoffman、M. Dinculescu 和 D. Eck，《音乐 Transformer：生成具有长期结构的音乐》，载于国际学习表征会议，2019 年。 [在线]。可访问：https://openreview.net/forum?id=4 7 rJe ShAcF X. Ma[2]、Y. Wang 和 Y. Wang，《基于图学习的偏好建模和文本查询的符号音乐生成》，IEEE 多媒体汇刊，第 26 卷，第 10 545–10 558 页，2024 年，doi：10.1109/TMM.2024.3408060。[3] Y.-S. Huang 和 Y.-H. Yang，《流行音乐 Transformer：富有表现力的流行钢琴作品的基于节拍的建模与生成》，载于第 28 届 ACM 多媒体国际会议论文集，2020 年，第 1180–1188 页，doi：10.1145/3394171.3413671。[4] S. Ji 和 X. Yang，《EmoMusicTV：利用层次化 Transformer 变分自编码器进行情感条件符号音乐生成》，IEEE 多媒体汇刊，第 26 卷，第 1076–1088 页，2024 年，doi：10.1109/TMM.2023. 3276177。[5] C. Raffel，《基于学习的序列比较方法及其在音频到 MIDI 对齐与匹配中的应用》，博士学位论文，哥伦比亚大学，2016 年。[6] B. L. Sturm，《GTZAN 音乐流派数据集分析》，载于 ACM 音乐信息检索用户中心多模态策略研讨会（MIRUM）论文集，2012 年，第 7–12 页，doi：10.1145/2390848.2390851。[7] E. Pampalk、A. Flexer 和 G. Widmer，《基于音频的音乐相似度和流派分类的改进》，载于国际音乐信息检索会议（ISMIR）论文集，2005 年，第 628–633 页。[8] F. Rodríguez-Algarra、B. L. Sturm 和 S. Dixon，《通过干预刻画音乐分类实验中的混淆效应》，国际音乐信息检索学会汇刊，第 2 卷，第 1 期，第 52–66 页，2019 年，doi：10.5334/tismir.24。[9] C. Hawthorne、A. Stasyuk、A. Roberts、I. Simon、C.-Z. A. Huang、S. Dieleman、E. Elsen、J. Engel 和 D. Eck，《利用 MAESTRO 数据集实现因子化钢琴音乐建模与生成》，载于国际学习表征会议，2019 年。 [在线]。可访问：https://openreview.net/forum?id=1 9 7r lYRjC F I. [10] Borovik、D. Gavrilev 和 V. Viro，《SyMuPe：富有情感且可控的符号音乐演奏》，载于 ACM 多媒体国际会议论文集，2025 年，第 10 699–10 708 页，doi：10.1145/3746027.3755871。[11] S. Tian、C. Zhang、W. Yuan、W. Tan 和 W. Zhu，《XMusic：迈向通用且可控的符号音乐生成框架》，IEEE 多媒体汇刊，第 27 卷，第 6857–6871 页，2025 年，doi：10.1109/TMM.2025.3590912。[12] L. Bradshaw 和 S. Colton，《Aria-MIDI：用于符号音乐建模的钢琴 MIDI 文件数据集》，载于国际学习表征会议，2025 年。 [在线]。可访问：https://openreview.net/forum?id=X5hrhgndxW [13] E. Choi、H. Kim、J. Ryu、J. Nam 和 D. Jeong，《关于 Lakh MIDI 数据集的去重》，载于第 26 届国际音乐信息检索学会会议论文集，2025 年，第 44–51 页，doi：10.5281/zenodo.17811316。[14] B. L. Sturm，《GTZAN 数据集：其内容、缺陷、缺陷对评估的影响及其未来使用》，2013 年，arXiv：1306.1461。[15] K. Lee、D. Ippolito、A. Nystrom、C. Zhang、D. Eck、C. Callison-Burch 和 N. Carlini，《对训练数据去重可提升语言模型性能》，载于第 60 届计算语言学协会年会论文集（第 1 卷：长篇论文），2022 年，第 8424–8445 页，doi：10.18653/v1/2022.acl-long.577。[16] S. Wu、D. Yu、X. Tan 和 M. Sun，《CLaMP：用于跨模态符号音乐信息检索的对比式语言-音乐预训练》，载于第 24 届国际音乐信息检索学会会议论文集，2023 年，第 157–165 页。[17] S. Wu、Z. Guo、R. Yuan、J. Jiang、S. Doh、G. Xia、J. Nam、X. Li、F. Yu 和 M. Sun，《CLaMP 3：跨越未对齐模态和未见语言的通用音乐信息检索》，载于计算语言学协会论文集，2025 年，第 2605–2625 页，doi：10.18653/v1/2025.findings-acl.133。[18] J. Serrà、E. Gómez、P. Herrera 和 X. Serra，“用于翻唱歌曲识别的色度二进制相似度与局部对齐”，《*IEEE* 音频、语音与语言处理汇刊》，第 16 卷，第 6 期，1138–1151 页，2008年，doi：10.1109/TASL.2008.924595。[19] X. Du、Z. Wang、X. Liang、H. Liang、B. Zhu 和 Z. Ma，“ByteCover3：针对短查询的准确翻唱歌曲识别”，载于 *IEEE* 国际声学、语音与信号处理会议，2023年，第 1–5 页，doi：10.1109/ICASSP49357.2023.10095389。[20] M. Douze、A. Guzhva、C. Deng、J. Johnson、G. Szilvasy、P.-E. Mazaré、M. Lomeli、L. Hosseini 和 H. Jégou，“Faiss 库”，2024年，arXiv:2401.08281。[21] Y. A. Malkov 和 D. A. Yashunin，“使用分层可导航小世界图进行高效且稳健的近似最近邻搜索”，《*IEEE* 模式分析与机器智能汇刊》，第 42 卷，第 4 期，824–836 页，2020年，doi: 10.1109/TPAMI.2018.2889473。[22] T. F. Smith 和 M. S. Waterman，“共有分子子序列的识别”，《分子生物学杂志》，第 147 卷，第 1 期，195–197 页，1981年，doi：10.1016/0022-2836(81)90087-5。[23] O. Gotoh，“一种改进的生物序列匹配算法”，《分子生物学杂志》，第 162 卷，第 3 期，705–708 页，1982年，doi: 10.1016/0022-2836(82)90398-9。[24] J. C. Platt，“支持向量机的概率”，载于《大间隔分类器进展》，A. J. Smola、P. L. Bartlett、B. Schölkopf 和 D. Schuurmans 编。麻省理工学院出版社，2000年，第 61–74 页，doi：10.7551/mitpress/1113.003.0008。[25] H. Zhang、J. Tang、S. R. M. Rafee、S. Dixon 和 G. Fazekas，“ATEPP：自动转录的表现性钢琴演奏数据集”，载于第 *23* 届国际音乐信息检索学会会议论文集，2022年，第 446–453 页。[26] F. Foscarin、A. McLeod、P. Rigaux、F. Jacquemard 和 M. Sakai，“ASAP：用于钢琴转录的对齐乐谱与演奏数据集”，载于第 *21* 届国际音乐信息检索学会会议论文集，2020年，第 534–541 页。[27] SecondHandSongs，“数据库简介”，作品与演奏元数据。 [在线]。网址：https://secondhandsongs.com/page/Introduction [28] W. Goebl，“维也纳 4×22 钢琴语料库”，维也纳音乐与表演艺术大学，数据集，2017年，doi: 10.21939/4X22。[29] Y. Kim、J. Park、J. Bae、K. Kim、T. Kwon、A. Lerch 和 J. Nam，“PianoVAM：多模态钢琴演奏数据集”，载于第 *26* 届国际音乐信息检索学会会议论文集，2025年，第 528–535 页。[30] Z. Wang、K. Chen、J. Jiang、Y. Zhang、M. Xu、S. Dai、X. Gu 和 G. Xia，“POP909：用于音乐编曲生成的流行歌曲数据集”，载于第 *21* 届国际音乐信息检索学会会议论文集，2020年，第 38–45 页。[31] P. Long、Z. Novack、T. Berg-Kirkpatrick 和 J. McAuley，“PDMX：用于符号音乐处理的大规模公有领域 MusicXML 数据集”，载于*IEEE* 国际声学、语音与信号处理会议，2025年，第 1–5 页，doi: 10.1109/ICASSP49660.2025.10890217。[32] K. J. M. Lee、J. Ens、S. Adkins、P. Sarmento、M. Barthet 和 P. Pasquier，“具备表现性音乐演奏检测特征的 GigaMIDI 数据集”，《国际音乐信息检索学会汇刊》，第 8 卷，第 1 期，1–19 页，2025年，doi: 10.5334/tismir.203。[33] J. Retkowski、J. St.epniak 和 M. Modrzejewski，“弗雷歇音乐距离：生成式符号音乐评估的一种度量”，2025年，arXiv:2412.07948。[34] DaoTwenty，“MIDI-GPT 的 PyTorch 实现”，GitHub 代码库。[在线]。网址：https://github.com/DaoTwenty/MIDI-GPT

附录 A 表示与配对特征

A. 解析、分词与检索

时值以 MIDI 四分音符单位计量，计算方式为用时钟刻度除以每四分音符脉冲数；不使用速度图，也不估算演奏拍数。若存在非打击乐音符，则排除第 10 通道音符；否则予以保留。同一通道、同一音高的音符开启事件按先进先出顺序关闭，最短时长为一个时钟刻度；未匹配的音符开启事件赋予0.25 个四分音符单位。不含音符的文件无效。舍入采用最接近的整数，遇到正好居中的情况时取偶数。全局起始时间箱为$\mathrm{round}(4t)$，并保留开头的静音部分。其最高和最低音高分别定义高音与低音轮廓；连续音程截断至 $[-36, 36]$。占用分箱之间的间隔Δt 得出节奏值$\min\{192, \max\{0,\mathrm{round}(12\Delta t)\}\}$。

高音/低音词元是音程 n-元组， $n = 2$, 3、4、5；节奏使用 $n = 2$、3、4。动机由连续三个高音音程/起音间隔对组成。对于 $C > 0$ 个可用起始位置，采样 0、 d、 $2d$、 $\ldots < C$，其中 $d =$最大值为$\{1, \lceil C/K \rceil\}$，且 $K = 2048$。和声使用仅基于起始位置的音级掩码，按 round(2*t*) 分箱；保留首个与最后一个非空分箱之间的空分箱，从首个分箱起最多保留 2048 个分箱。长度为 2、4 和 8 的掩码序列，会在 12 种常见音级旋转下规范化为字典序最小的形式，每种长度取 $K = 512$ 个起始位置。

每种词元族都使用 8192 个分箱中的非负 FeatureHasher 计数，权重为 (1+log c){1+log[(N+1)/(df+1)]}，并采用 ℓ_2 归一化。逆文档频率在检索语料库内拟合； N 包含无效的零行。音程直方图有 75 个分箱，按 $\delta + 37$ 编索引。时长/起音间隔直方图有 16 个分箱，按

$$\mathrm{clip}\big(\mathrm{round}[2\log_2 \max(t, 1/32)] + 8, 0, 15\big).$$

时长计数涵盖所有音符，起音间隔计数则基于连续的非空分箱。全曲色度权重将音级 $p \bmod 12$ 乘以 max(d, 0。125)max(v, 1)，并考虑时长 d 和力度 v。直方图会 ℓ_2 归一化；零向量保持为零。

Token 检索向量使用稀疏随机投影，投影到 min$\{256,$ max(8, $N_{\text{valid}} - 1)\}$ 维，然后以 FP32 格式存储并重新归一化。随机种子为 $20260911 + v$，视图顺序依次为低音、和声、动机、音程、时长；直方图视图直接建立索引。每个 FAISS IndexHNSWFlat 使用内积、 $M = 32$、efConstruction$= 160$ 和 efSearch$= 200$，插入顺序固定且使用单线程。最多查询 101 个条目，移除自身，并保留至多 100 个邻居。有效文件少于两个时，不产生候选项。公式 (6) 应用于同一视图的互惠 Top-50 筛选。全局特征使用未经投影的向量。

B. 57 个坐标与边界情况

对于高音旋律线、低音、节奏、和声和动机，每组前 15 个坐标依次表示余弦相似度、杰卡德相似度和包含度。对于占用分箱集合 S_A、 S_B 和$I = |S_A \cap S_B|$，后两项为

$$J = \frac{I}{\max(|S_A \cup S_B|, 1)}, \quad K = \frac{I}{\max(\min(|S_A|, |S_B|), 1)}.$$

接下来是音程、时长和起音间隔的余弦相似度，然后是对 12 种移位取最大值的全局音级余弦相似度。其余十个全局坐标依次比较音符数、占用起音分箱数、最终偏移量、音轨数、音高最小值/最大值/均值/总体标准差、不同乐器音色数，以及占用的半四分音符分箱。音高统计量采用绝对差值；其他标量采用$\min(a, b)/\max(a, b, 10^{-6})$。音轨数的计算沿用上述打击乐回退方案。

接下来为低音、和声和动机附加（R_v， M_v）：

$$R_v = \frac{1}{\min(r^v_{AB}, r^v_{BA})}, \quad M_v = \mathbf{1}[\max(r^v_{AB}, r^v_{BA}) \le 100].$$

等级超过 100 或缺少邻居时取值为 101。这六个坐标位于局部区块之前； $\sum_v M_v$ 和 max$_v$ R_v构成向量末尾，使检索坐标总数达到八个。

对于保留路径 $\ell = 1$, ..., P，令 Q_ℓ， m_ℓ， U_ℓ， V_ℓ分别表示得分、匹配数和匹配索引集合，且 $M = \sum_\ell m_\ell$。定义

$$c_A = \frac{|\cup_\ell U_\ell|}{n_A}, \qquad c_B = \frac{|\cup_\ell V_\ell|}{n_B},$$
$$C = \frac{2c_A c_B}{\max(c_A + c_B, 10^{-8})}.$$

合并跨度为 $T_\ell = \max U_\ell - \min U_\ell + \max\ V_\ell - \min V_\ell + 2$。匹配加权间隙比例为

$$G = \sum_\ell \frac{m_\ell}{M} \frac{\max(0, T_\ell - 2m_\ell)}{T_\ell}.$$

14 个路径坐标依次为： $S = \max_\ell (Q_\ell/m_\ell)$ 、$\sum_\ell Q_\ell/M$、max$_\ell\, m_\ell$、 M、 P、短序列/长序列覆盖率、 C、 G，以及式（7）中的五个匹配对平均相似度（a_m， a_b， a_r， a_h， a_d）。覆盖率使用保留序列长度；若长度相同，则将A 分配给短序列项。

¯对于归一化片段音级特征 h_{Au}，令 h_A 为归一化和 $\sum_u h_{Au}$，并定义 $t_s = \bar{h}_A^{\mathsf{T}}$ roll($\bar{h}_B, s$)。随后附加 t_0 和 $t_0 - \max_{s\neq 0} t_s$，再加上四个交互项：min(a_m, a_b, a_r, a_h)、($a_m + a_b + a_r + a_h$)/4、 SC，以及$S(1 - G)$。因此，局部区块包含18个由路径导出的坐标和两个独立音级坐标。对于存在有效序列但没有被接受路径的情况，仅保留两个音级值；提取失败会将全部20个局部坐标置零。零向量/空集合会产生零相似度。特征采用 FP32，并将 NaN/$+\infty$/$-\infty$ 替换为 0/1/0。规范输入顺序 $A < B$固定对齐方向；返回的 HNSW 顺序固定排名并列。

附录 B 对齐、学习与分组

A. 局部序列与回溯

减去最早的起始后，保留n =最小值$\{512,$最大值(1, $\lceil T/2 \rceil)\}$ 个由两个 MIDI 四分音符单位组成的分段，其中 T 是以相同单位表示的最终偏移量 。位于 t的音符进入分段 $u = \lfloor t/2 \rfloor$ 和分箱（$\{7, \lfloor 4 \rangle \rfloor\}$；丢弃 $u \ge n$，并保留内部的空分段。每个分箱记录最高/最低起始音高以及所有起始次数。色度在音符的音级处添加 最小值(d, 2) 最大值(v, 1)，且完全位于其起始分段内。分段

色度和起始计数向量分别进行归一化。音高相似度对同时有效的分箱取平均；不存在同时有效的分箱时，结果为零。两个空分段的音高、节奏和和声相似度均为零，但密度相似度为 1。

在零移调且边界状态为零时应用 (8)。按行、再按列扫描，保留第一个最大值。正状态并列时优先选择对角线、E，再选择 F；间隙并列时优先选择开启而非延伸。回溯在零值重启或到达边界时停止，并输出对角线匹配。每接受一条路径后，将奖励矩阵中与其匹配的行和列设为 -10^6；后续路径只能通过间隙跨越这些位置。达到三条路径、最大值非正，或最佳路径的匹配数少于四个时停止，即使存在得分较低但符合条件的路径也不例外。

B. 工作不相交拟合与校准

LMD 开发队列包含 16,788 个校验和唯一文件和 10,098 个作品，每个录音标识符对应一个文件。使用每个作品标识符的 SHA-256 哈希值前八位十六进制数字对 100 取模。桶 $[0, 70)$、$[70, 85)$ 和$[85, 100)$ 分别包含 11,740/7,049、2,501/1,500 和 2,547/1,549 个文件/作品，用于拟合、校准和测试。拟合使用连续的五个 14 桶分片。每个分片分别进行逆文档频率计算、索引和候选选择；将其特征行拼接起来用于训练。校准/测试池独立构建。正样本属于同一作品；负样本属于不同作品。录音/校验和唯一性是开发阶段的限制，而非推理筛选条件。

对于种子 20260911–20260913，保留全部 8,634 个正样本候选项，并在不放回的情况下为每个正样本最多抽取 50 个负样本，从而为每个分类器得到 440,334 个样本。将样本打乱，并拟合类别均衡的直方图梯度提升分类器模型：迭代 260 次，学习率为 0.05，最多 15 个叶节点，叶节点最小大小为 80，ℓ_2 强度为 8，且不进行提前停止（scikit-learn 1.7.2；NumPy 1.26.4）。

在完整校准池（1,800 个正样本）上，对均值分类器概率拟合逻辑回归，使用 lbfgs、$C = 1$，且最多迭代 1,000 次。拟合值为 a =9。548566767981713 和 $b_0 = -9.098677121144501$；将 Sigmoid 参数截断至 $[-50, 50]$。按降序评估不同得分，并接受完整的并列组。要求精确率 ≥ 0.97 且至少接受 20 对，然后按 (F1 值、召回率、精确率) 进行字典序最大化；剩余并列时按阈值降序打破平局。若不可行，则不返回阈值。冻结的工作点为 $\tau = 0.5975961974276196$。表示/候选变更需要重新拟合分类器、校准和阈值。

C. 隔离与独立语料审计器

初始化单例组，并将每条达到阈值的正向边进行并集操作。边的顺序不会影响连通分量；不存在大小上限、得分排序或桥接计数约束。使用固定的验证集/测试集应用（10）。无效文件仍作为未解决的单例保留。错误桥接可能移除符合条件的内容，而遗漏的关系可能造成暴露。

表V校准后的 F1 值（%）及基于工作自助法的 95%置信区间。并集：CAugBERT ∪CLaMP。完整校准结果：请参阅项目页面上的扩展结果。

数据集	MVA-SWI	95% 置信区间	并集
SHS	72.99	[51.28, 85.71]	39.68
ASAP	99.73	[99.40, 99.93]	60.40
ATEPP	98.24	[97.15，99.09]	68.20
LMD-clean	82.35	[78.77，85.81]	51.12
维也纳 4×22	100.00	[100.00，100.00]	90.92
PianoVAM	89.17	[71.50，95.09]	57.33

这七个语料库审计器采用相同的检索—对齐结构，但要求至少在两个视图中获得相互 Top-20 支持、通过一个 29 特征的初步门控，并使用一个独立训练的 49特征验证器，分数阈值为 0.9396。该配置在各语料库之间保持固定，并且不同于经过校准的 57 特征核心检测器。其更严格的检索/门控可能会遗漏核心候选项；报告的暴露量以其推断组为条件。部署重放按得分降序处理边，要求较大的合并包含多个正向链接，并限制组大小。发布的部署配置记录了这些分组参数。

A附录 C实验方案和关键结果

A. 队列、校准和敏感性

成对队列包含的文件/符合条件的正例：SHS 306/80、ASAP 1,067/4,988、ATEPP 1,527/3,339、LMD-clean 2,547/2,037、维也纳 4×22 88/924，以及 PianoVAM94/77。正样本共享同一作品，但录音 ID 不同；预测采用相同的排除规则。召回率包含检索遗漏。表V报告了独立校准的 F1 值。其 95% 置信区间采用 2,000 次作品分组自助法重复抽样（种子 20260921）计算，汇总查询级 TP/FP/FN；每个无序对在重采样前计入两次。LMD-clean 是附录 B-B 中按作品划分的测试队列。

特征消融在相同精确率约束下，使用 20260911–20260913 的种子重新训练/校准，并固定候选项；完整/无全局/无检索/无局部维度分别为 57/28/49/37。平均精度使用已评分的候选项计算。敏感性取 LMD-clean、ATEPP 和 ASAP 的均值。在每个视图数量下，MAESTRO 验证集依据检索到的正样本数最多、候选项数最少、视图字典序依次选择子集；该子集保持不变地迁移使用。Top-k 使用 $k = 10$、25、50、75、100 的全部视图。两者都要求同视图互惠支持，并排除同一录音配对。集成规模 1–5 分别重新校准种子 20260911–20260915 的全部 5/10/10/5/1 个子集。均值/总体标准差总结了相互重叠的子集；三分类器均值不同于固定的三种子消融。

B. 对齐与时序检查

移位/前缀测试保持候选项和全部 57 个特征固定（扩展结果）。在一项单独构造的两跳测试中，共享位置过滤保留了 54/54 个真实链接，且没有错误链接（0/174）；而位置打乱后，则保留了 29/54 个真实链接，并产生了 103/174 个错误链接。

这不属于固定分组，也没有在特征相同的情况下单独比较有序匹配与无序匹配。仅后段乐句和乐句重排序尚未评估。

ASAP 的节奏分层采用标注的拍间间隔中位数的较大值/较小值；ATEPP 则采用音符跨度时长，该指标也能反映内容（扩展结果）。这些标注仅用于诊断；分层中包含检索遗漏，且无法衡量局部速度自由变化。在候选项、分类器和阈值固定的情况下，我们将两种间隔惩罚均缩放为 0.5/1.5，测试权重相等以及强调旋律/和声的权重，并在输入覆盖范围固定时采用 1/2/4 个四分音符的片段。在 ASAP、ATEPP 和 LMD-clean 上，F1 值变化最多为 0.6 个百分点，精确率/召回率的变化均不到一个百分点。权重相等与默认权重表现相近；其他方案未单独优化。

C. 独立标签与暴露控制

独立标签依据 ATEPP 作品标识、ASAP乐谱目录、LMD-clean 艺术家/歌曲分组、MAESTRO 作曲家/标题、维也纳乐谱标识、PianoVAM 作品元数据以及 SHS 翻唱作品分组 [9]，[13]，[25]–[29] 确定。演奏、转录、移调、编曲和无歧义片段共享一个作品标签；乐章、混曲和引用片段则单独标记。仅凭标题/风格/动机相似性不足以确定标签。两位审阅者最初分别确认了所选 100 对中的 98 对和 99 对，并将其余配对标记为不确定。共同审阅后，100 对均得到确认，其中包括六对作品–片段配对。

在这项受控研究中，Clean 排除了选定的测试集/供体组件及文件身份。无关/相同组以匹配词元替换相同窗口。先在每个组内计算配对训练种子的损失均值，再计算该组的相对降幅；各组权重相同。表 I 的置信区间采用 10,000 次分组自助法重复样本，固定训练种子，不计入选择和新训练的不确定性。另行重新训练的 5k/10k 运行保留 20k 学习率调度，但使用 RTX 3090/4090 而非 RTX 5090，因此无法将结果单独归因于更新次数（见扩展结果）。ATEPP、ASAP 和 LMD-clean 的清理结果是对 20 组配对的 80/10/10 划分取平均；仅移除训练文件。清理耗时不包括检测。

D. 语料库审计与阈值迁移

表六独立验证了采用 49 个特征的 Fast 审计者与采用 57 个特征的核心版本。七语料库重放使用单独的检查点/分组配置。在 200 次随机 80/10/10 划分中，百分位区间描述的是以推断组为条件的划分变异性；官方划分则使用清单元数据。遗漏关系可能会降低估计的暴露程度，而错误桥接则可能使其升高。

两位审阅者对每个语料库中的 100 条随机边、20 条定向弱桥接以及按规模选取的五个连通分量进行了评估协调（表 VII）。定向抽样用于描述失败模式，而非估算其普遍程度；未标注语料库中的版本漏检率仍然未知。检索覆盖率、连通分量成员关系、解析计数和运行时间见扩展结果。4,352 个解析失败项仍是未解决的单例。

表六Fast Auditor 与 v1.9 核心的独立验证。成对精确率、召回率和 F1 值均以百分比表示；召回率包含检索遗漏。

队列	模型	P	R	F1
ATEPP	Fast	98.37	97.36	97.86
	v1.9	98.08	97.72	97.90
LMD-clean	Fast	91.70	62.89	74.61
	v1.9	96.18	67.99	79.67
ASAP	Fast	99.70	99.66	99.68
	v1.9	99.84	99.62	99.73

表 VII未标注语料库的人工审计。边/桥报告同一作品/不同作品/不确定计数。混合组件包含多个已确认作品。

语料库	随机边	定向桥	混合/已审计
PDMX	95/3/2	17/2/1	2/5
LMD-full	91/7/2	15/4/1	3/5
Aria-MIDI	93/4/3	16/3/1	2/5
GigaMIDI	94/5/1	16/4/0	2/5

表 VIII原始五种条件的训练池。暴露计数统计具有训练版本的测试作品；保留率为干净令牌保留率（%）。并集为默认嵌入并集。

条件	文件	暴露	保留率
	ATEPP		
未校正	1223	121/121	100.00
随机移除	486	72/121	63.97
MVA-SWI	486	0/121	98.91
并集	209	18/121	30.88
标签参考	495	0/121	100.00
	LMD-clean		
未校正	1223	73/142	100.00
随机移除	1023	63/142	85.19
MVA-SWI	1023	25/142	99.67
并集	1107	50/142	99.53
标签参考	965	0/142	100.00

开发集选定的阈值在留出数据源上固定不变。没有任何经过测试的核心设置达到 0% 目标。LMD-Clean 的 1%/5% 目标分别选用 0.5/0.605 的阈值，对应的暴露率为 12.95%/20.55%，保留率为 92.63%/99.24%。在 5% 目标下，ATEPP/ASAP/MAESTRO 的暴露率分别为 0.36%/0%/0%，保留率分别为 45.74%/44.53%/59.89%。因此，开发集目标未必能迁移。我们分发汇总结果；原始文件仍受上游许可证和第三方权利约束。

E. 下游任务与预算匹配队列 *s*

表 VIII 给出了原始的五个下游数据池；图 6 还额外纳入了校准并集。每种条件/模型/数据集均使用三个配对随机种子，并采用固定的评估内容。只有随机移除和 MVA-SWI 的文件数量相匹配。

单独的 ATEPP研究（表 IX）在三个种子、20k 次更新、11,316 个窗口和 5,884,369 个计分词元上，匹配了 MIDI-GPT 的初始权重。独立参考

表 IX预算匹配的 ATEPP 结果。偏差为相对于独立参考的每部作品平均有符号负对数似然偏差。配对 95% 置信区间以训练种子为条件。

条件	暴露 作品数	Dev. (%)	95% 置信区间（%）
未校正	119/121	−21.77	[−22.93, −20.59]
精确去重	119/121	−21.77	[−22.93, −20.59]
随机移除	72/121	−10.11	[−11.31, −8.93]
MVA-SWI	0/121	+2.15	[+1.85, +2.48]
原始参考	0/121	+0.89	[+0.63, +1.19]
独立参考	0/121	0.00	[0.00, 0.00]

使用标签排除验证集/测试集作品，并在固定配额下通过确定性哈希选择窗口。原始参考在相同文件预算下修复了MVA-SWI池。评估固定为 121 个作品、152 个文件、1,192 个窗口和 602,471 个词元。先在文件内对有效词元的负对数似然取平均，再对作品内的文件取平均，最后对种子取平均。配对作品自助法置信区间以训练种子为条件。

图6使用了1,223个未校正ATEPP文件，涉及121/121个暴露作品；图7则使用1,140个文件，涉及119/121个。不同的数据池、预算和参考项构成了各研究的区别；其绝对偏差/有符号偏差不可互换。前者报告配对种子标准差，后者报告作品自助法置信区间。

附录D 听觉与生成评估

A. 生成器训练与解码

模型从头开始训练，使用包含237个词元的REMI风格词汇表和1,024个词元的上下文：每个 $4/4$小节16个位置、128种音高、64种截断时值、16个力度分箱、八组有音高的乐器和鼓组，以及小节/序列控制。LMD使用四小节提示，总长度最多为16小节。最终检查点在教师强制下对有效的续接词元进行评分。

MIDI-GPT [34] 使用6个块、512的宽度、8个注意力头，参数量为19.68M。作为参照的Transformer包含10个预归一化因果块，宽度为384、6个注意力头、前馈层宽度为1,536、dropout为0.1、旋转位置编码以及权重绑定的嵌入。长短期记忆网络使用两层、每层512个单元，嵌入宽度为256，dropout为0.1；TCN使用9个残差块，宽度为384、前馈层宽度为1,536、卷积核为5，膨胀率为 2^0–2^8。变分自编码器包含2层编码器和3层解码器，宽度为128、4个注意力头、潜变量宽度为64，KL权重为0.01。扩散模型使用潜变量/模型宽度64/256和1,000个线性噪声步数（10^{-4}–0.02）。设置见configs/；Transformer-S/M/L架构见checkpoint model_config。

受控 LMD 训练使用 AdamW（$\beta_1 = 0.9$, β_2 =0.95）、批量大小 32/微批量大小 8、梯度裁剪值 1.0、1,000 个预热步数，并采用余弦衰减至 20k 个更新步数。MIDI- GPT/LSTM/参考 Transformer 使用学习率 2×10^{-4}，衰减率为 0.1；变分自编码器/扩散使用 3×10^{-4}/5×10^{-4}，衰减率为 0.01。解码使用 128 个提示词元和 512 个新词元，温度为 1.0，top-$p = 0.95$。

表X长段落匹配率（%，均值 ±标准差，基于十个训练随机种子）。这些指标衡量的是与留出续写的一致性，而非感知质量。

条件	MIDI-GPT	Transformer
ATEPP		
未校正	22.42 ± 0.92	23.89 ± 0.71
随机移除	10.92 ± 0.94	11.69 ± 0.42
MVA-SWI	3.64 ± 0.27	4.06 ± 0.46
默认并集	2.28 ± 0.10	3.03 ± 0.29
标签参考	3.33 ± 0.17	3.94 ± 0.43
LMD-clean		
未校正	21.90 ± 1.86	22.30 ± 1.40
随机移除	10.48 ± 1.37	11.01 ± 0.22
MVA-SWI	2.46 ± 0.56	3.33 ± 0.78
默认并集	1.29 ± 0.26	2.13 ± 0.70
标签参考	2.41 ± 0.49	3.22 ± 0.90

B. 盲评与不确定性

聆听评估涵盖五种条件、两个模型/数据集和十个训练种子，不包括校准并集。97 名招募参与者中，有 13 名未完成答题，另有 4 名至少未通过两项注意力检查，因此予以排除。

保留的 80 名听众按音乐背景平均分组。每人评审 12 组各含 5 个片段的内容、2 次盲重复和 3 项注意力检查。96 组共包含 480 个片段：来自 12 条清理测试提示，每种条件/数据集/模型对应 24 个片段。每个片段获得 10 次评分，每个背景组各 5 次，总计 4,800 次正式评估；800 次重复评分不计入。匿名标签经过随机排序。在安静环境中，通过耳机播放以每分钟 120 拍/16 千赫渲染的 8 秒片段，并完整播放；允许重播。

评分涵盖整体音乐性、旋律、节奏和和声（1–5）。“无法判断”视为缺失：19,200 个回答中有 18,347 个有效，其中旋律/节奏/和声评分缺失分别为 222/24/607 个，而整体评分完整。均值使用可用回答计算；对比使用匹配的非缺失回答。先在每个种子内取均值，再对各种子等权加权。使用 10,000 个交叉的听众–作品bootstrap 重采样，固定训练种子。整体音乐性和 MVA-SWI 减去未校正值是主要指标；其余三个整体对比使用同时 95% 置信区间。次要维度仅作描述性分析。在每种设置下，四个整体对比均包含零；未预先规定等效界值。

C. 自动指标

评估固定使用 120 个 ATEPP 作品/119 个 LMD-clean 分组、五个条件、两个模型、十个种子，以及每个分组/检查点十个生成结果：共 239,000 个片段。表X在种子内汇总生成结果，并报告种子均值/标准差。匹配要求在四拍内至少包含 16 个音符，并依据固定规则，音高和起音间隔差异均不得超过两个。完整的旋律/节奏相似度和 FMD-CLaMP 1 [16], [33] 结果及定义见扩展结果。