Musicology-Guided Structural Attention for Symbolic Classical Music Form Analysis

讯谱 Studio 曲式分析模块：程序完善与论文研发路线

目标顺序：先把程序做成可训练、可评测、可复现实验系统，再进入论文写作。

项目： Xunpu Studio / 讯谱 Studio

研究主题： Symbolic Score Form Analysis

核心机制： Musicology-Guided Structural Attention (MGSA)

文档版本： Research Plan v1.0

日期： 2026-10-05

目录

1 项目定位与当前结论 **2**

2 现有原型：已经具备什么 **2**

3 核心模型：MGSA **3**

3.1 总公式 . . . 3

3.2 为什么 boundary 应当是负项 . . . 3

4 六类音乐学先验的可实现定义 **3**

5 推荐的层级架构 **4**

5.1 输入表示 . . . 5

5.2 Measure Pooling 改造 . . . 5

6 程序完善顺序：先做工程闭环 **6**

6.1 Phase 0：建立一个绝对可靠的 baseline . . . 6

6.2 Phase 1：统一数据结构 . . . 6

6.3 Phase 2：六类 prior builder . . . 6

6.4 Phase 3：MGSA 模块工程化 . . . 7

6.5 Phase 4：多任务输出 . . . 7

7 数据集策略 **7**

7.1 S3：最适合验证“音乐学先验” . . . 7

7.2 Classical Form Benchmark：扩大 form/phrase 监督 . . . 8

8 评测指标与实验设计 **8**

8.1 核心指标 . . . 8

8.2 必须做的 baseline . . . 8

9 消融实验矩阵 **9**

10 论文创新边界与相关工作 **9**

11 论文结构：程序跑通后再写 **10**

11.1 论文主结果表预留模板 . . . 11

12 推荐代码仓库结构 **11**

13 近期执行清单 **12**

14 建议的第一版研究题目 **13**

15 参考文献与核验入口 **13**

1 项目定位与当前结论

本项目不是简单地把乐谱 token 输入 Transformer，再预测 A/B/A′ 标签，而是尝试构建一个**面向音乐结构认知的符号乐谱分析器**。系统的目标是从 MusicXML 或其他结构化乐谱表示出发，在小节、乐句、段落和整曲多个层级上同时建模局部凝聚、结构边界和远距离主题复现。

核心研究问题

给定符号乐谱，能否把节拍、动机、和声、终止、重复和边界等音乐学先验直接注入注意力计算，使模型比普通 Transformer 更准确地识别乐句边界、段落边界、A/A′ 关系及整体曲式?

当前最合适的论文表述不是“发明一种全新的音乐注意力”，而是:

建议的研究表述

We propose a musicology-guided structural attention mechanism for hierarchical form analysis of symbolic classical music. The mechanism injects explicit metric, motivic, harmonic, cadential, repetition and boundary priors into attention logits, and jointly optimizes boundary detection and section-relation modeling.

这一定义把创新点收敛到三个可验证部分:

- **对象明确:** symbolic score，而不是通用音频结构分析;
- **机制明确:** priors 直接作用于 attention logits，而不是仅作为输入 feature;
- **任务明确:** hierarchical form analysis, 包括 boundary、section relation 与 form decoding。

2 现有原型：已经具备什么

当前原型已经包含两层关键代码结构：`FormAwareScoreAttention` 与 `XunpuFormModel`。前者已经实现多种先验偏置的加权注入，并对 boundary 施加负向抑制；后者已经包含小节聚合、结构注意力、边界预测、self-similarity matrix 以及 Same/Variation/Contrast 段落关系分类的雏形。

当前原型最有价值的部分

不是“有一个 Transformer”，而是已经把 **音乐学吸引项**与 **结构边界抑制项**写进了同一个注意力公式。这为后续消融实验留下了天然接口。

当前原型仍属于 **research prototype**，不能直接当作已验证模型，原因包括:

- 先验 bias 目前主要是接口与模拟张量，尚未形成稳定、统一的数据提取管线;
- measure pooling 仍是演示式实现，效率和语义表达都需要改进;
- boundary target、section relation target、form label 尚未形成规范数据格式;

- 没有训练日志、baseline、ablation、cross-validation 和独立测试集结果；
- 没有验证各先验之间是否互相冗余，或者某些先验是否引入标签泄漏。

3 核心模型：MGSA

3.1 总公式

建议将机制正式命名为：**Musicology-Guided Structural Attention (MGSA)**。核心注意力得分定义为：

$$S_{ij}^{(h)} = \frac{Q_i^{(h)} K_j^{(h)\top}}{\sqrt{d_h}} + \lambda_m B_{ij}^{metric} + \lambda_o B_{ij}^{motif} + \lambda_h B_{ij}^{harmony} + \lambda_c B_{ij}^{cadence} + \lambda_r B_{ij}^{repetition} - \lambda_b B_{ij}^{boundary}. \quad (1)$$

归一化后：

$$A_{ij}^{(h)} = \mathrm{softmax}_j\left(S_{ij}^{(h)}\right), \qquad H_i^{(h)} = \sum_j A_{ij}^{(h)} V_j^{(h)}. \quad (2)$$

其中 $B^{metric}, B^{motif}, B^{harmony}, B^{cadence}, B^{repetition}$ 为结构吸引项，$B^{boundary}$ 为结构抑制项。所有 λ 建议使用非负参数化，例如 $\mathrm{softplus}(\theta)$。

3.2 为什么 boundary 应当是负项

普通 self-attention 倾向于在全序列中自由建立联系，但音乐曲式同时包含两种相反约束：

1. 同一乐句或同一段落内部应当保持较强的局部凝聚；
2. 跨越很远距离的主题再现（例如 A 到 A’）又必须允许建立联系。

因此模型需要同时具有：

$$\text{Local Structural Cohesion} \quad \leftrightarrow \quad \text{Long-range Thematic Recurrence}. \quad (3)$$

$-\lambda_b B^{boundary}$ 用来阻止注意力在明确结构边界处无条件泄漏，$+\lambda_r B^{repetition}$ 则为远距离主题复现提供“跨边界通道”。这组竞争关系是后续论文最值得验证的机制之一。

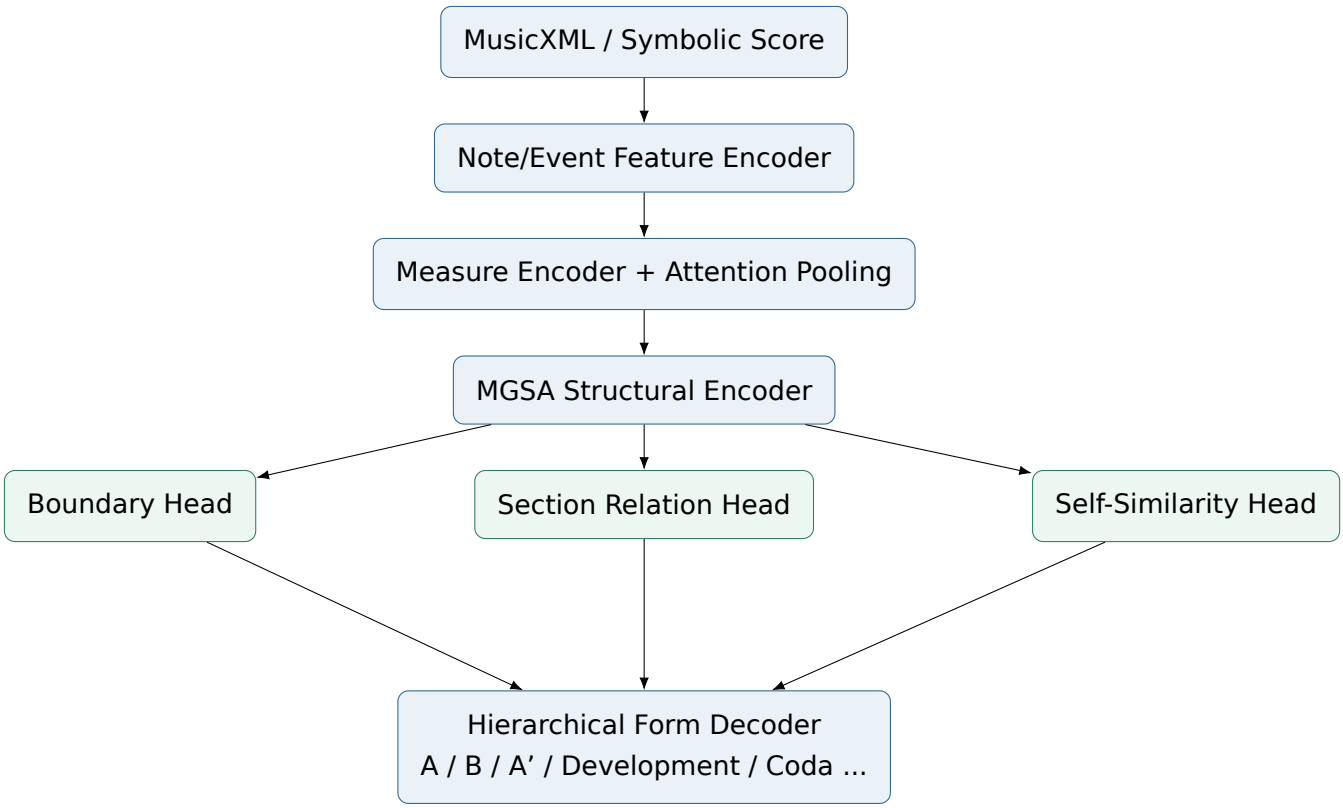
4 六类音乐学先验的可实现定义

Bias	音乐学含义	程序实现建议
B^{metric}	节拍与超小节位置关系	beat strength、downbeat、measure index modulo phrase length；先从强拍/小节首相似矩阵开始。
B^{motif}	动机、音程轮廓和节奏型相似	pitch interval sequence + rhythmic ratio；使用局部 n-gram embedding 或 DTW/cosine similarity。
$B^{harmony}$	调性、和弦功能与 harmonic proximity	key、Roman numeral、chord root/quality；先使用功能类别 embedding，再考虑 Tonnetz/tonal distance。
$B^{cadence}$	终止式与乐句闭合证据	cadence detector 输出 PAC/IAC/HC/DC 或更粗粒度 cadence confidence；作为位置级和 pairwise bias。
$B^{repetition}$	跨段重复、变奏和再现	measure/phrase embedding self-similarity；允许移调不变、节奏轻微变化和配器变化。
$B^{boundary}$	乐句或段落边界抑制	由 cadence、novelty、double bar、key change、texture shift 等产生 boundary probability，再转换为跨区间 penalty matrix。

避免标签泄漏

如果 ground-truth form boundary 直接被用于构造 $B^{boundary}$，模型就会提前看到答案。训练和测试时只能使用**可自动观测或由独立模块预测**的边界证据，例如终止概率、调性变化、复纵线、novelty 等；人工标注 boundary 只能用于监督 target。

5 推荐的层级架构



5.1 输入表示

第一版不要把输入做得过度复杂。建议每个 note/event 至少包含：

- pitch class、MIDI pitch、interval from previous note；
- duration、onset、beat position、measure position；
- voice/instrument；
- key estimate、chord class（如果可靠）；
- articulation/dynamics 可先保留为 optional feature。

音符先编码为 e_i，再聚合为 measure representation m_t。第一阶段论文实验建议以 **measure 为主序列单位**，因为曲式边界与数据标注通常更容易对齐到小节级，且显著降低长序列计算成本。

5.2 Measure Pooling 改造

当前逐小节 Python for-loop 的演示式 pooling 应替换为向量化实现。推荐顺序：

1. masked mean pooling 作为最稳定 baseline；
2. learnable attention pooling；
3. 如有必要再测试 Set Transformer 或 hierarchical Transformer pooling。

不要一开始就把 pooling 做得过于复杂，否则论文难以判断性能提升究竟来自 MGSA 还是聚合器。

6 程序完善顺序：先做工程闭环

6.1 Phase 0：建立一个绝对可靠的 baseline

先实现普通 Transformer Encoder，不加任何音乐学 bias：

```
MusicXML -> feature extractor -> measure embeddings
       -> Transformer encoder
       -> boundary head / relation head
```

这一步的意义不是追求最好结果，而是提供论文后续所有比较的基准。没有 baseline，MGSA 的“提升”无法成立。

6.2 Phase 1：统一数据结构

建议定义统一样本 JSON：

```
{
  "piece_id": "...",
  "measures": [...],
  "note_features": [...],
  "measure_index": [...],
  "boundary_targets": [...],
  "phrase_spans": [...],
  "section_spans": [...],
  "section_labels": [...],
  "relation_pairs": [...],
  "form_label": "..."
}
```

训练前把 MusicXML 完全转换为缓存数据，避免训练 loop 中实时跑复杂 music21 分析。

6.3 Phase 2：六类 prior builder

所有先验统一成：

$$B^{(k)} \in \mathbb{R}^{T\times T}, \qquad B^{(k)}_{ij} \in [0,1] \tag{4}$$

并提供同一接口：

```
prior_builder(score) -> {
   "metric": [T, T],
   "motif": [T, T],
   "harmony": [T, T],
   "cadence": [T, T],
```

```
    "repetition": [T, T],
    "boundary": [T, T]
}
```

每种 prior 必须可以单独开关，以服务后续 ablation。

6.4 Phase 3：MGSA 模块工程化

程序层面至少补齐：

- padding mask 与 variable-length score；
- per-head 或 shared λ 两种模式；
- residual + pre-norm Transformer block；
- dropout、attention map export；
- bias normalization/clipping，防止某一 prior 数值范围压倒 QK^T；
- mixed precision 与 batch inference。

6.5 Phase 4：多任务输出

建议第一篇论文只保留三个核心任务：

1. **Boundary Detection**：小节 t 后是否形成结构边界；
2. **Section Relation**：Same / Variation / Contrast；
3. **Global Form Classification 或 Form Decoding**：作为高层结果。

总损失：

$$\mathcal{L} = \alpha \mathcal{L}_{boundary} + \beta \mathcal{L}_{relation} + \gamma \mathcal{L}_{form} + \eta \mathcal{L}_{contrastive}. \tag{5}$$

其中 contrastive loss 用于约束 $D(A, A') < D(A, B)$，可以后加，不要让第一版训练系统过度复杂。

7 数据集策略

7.1 S3：最适合验证“音乐学先验”

S3 (Symbolic Symphony Set) 提供 4 部交响曲、16 个乐章，并包含 MusicXML 及 form analysis、phrase、harmony、cadence 和 orchestral texture 标注。它与本项目的先验设计高度匹配，因此非常适合作为**机制验证集与可解释性展示集**。

但 S3 规模较小，不应单独承担“模型泛化能力”的全部证据。更合适的用途是：

- 验证 cadence/harmony/boundary prior 是否合理;
- 画 attention heatmap 与 SSM;
- 做 case study;
- 做小规模 cross-validation。

7.2 Classical Form Benchmark：扩大 form/phrase 监督

Szelogowski 等人在 ISMIR 2022 发布了 200 首西方古典音乐作品的 form 与 phrase 标注数据，并给出了 form classification 与 phrase analysis/classification benchmark。该资源适合承担更大规模的结构识别对比。

数据组合原则

不要把不同数据集的 label ontology 强行合并。例如“phrase / theme / section / exposition / chorus”等标签层级不同。第一版系统应建立清晰的 label mapping，并把无法严格对齐的任务分开评测。

8 评测指标与实验设计

8.1 核心指标

任务	建议指标
Boundary Detection	Precision / Recall / F1；同时报告 strict measure-level 与允许 ± 1 measure tolerance 的结果。
Section Relation	Accuracy + Macro-F1；单独报告 Same、Variation、Contrast 三类。
Form Classification	Accuracy + Macro-F1；类别不均衡时必须报告每类结果。
Segmentation Quality	可补充 pairwise/grouping 指标；如果与音频 MSA 文献比较，应明确时间单位从 frame 转为 measure 后的差异。
Interpretability	attention 与人工 cadence/boundary/repetition annotation 的相关性或 retrieval precision。

8.2 必须做的 baseline

至少包含:

1. BiLSTM 或简单 sequence baseline；
2. Vanilla Transformer；
3. Vanilla Transformer + raw musicological features（仅拼接输入，不改 attention）；
4. MGSA full model。

第三个 baseline 非常重要。它能回答一个审稿人必问的问题：

关键实验问题

性能提升究竟来自“模型看到了更多音乐学 feature”，还是来自“把这些 feature 作为 pairwise structural bias 注入 attention 的方式”本身？

9 消融实验矩阵

Model	Metric	Motif	Harmony	Cadence	Repeat	Boundary	Purpose
Vanilla	-	-	-	-	-	-	基线
+ Metric	+	-	-	-	-	-	节拍先验贡献
+ Motif	-	+	-	-	-	-	动机相似贡献
+ Harmony	-	-	+	-	-	-	和声贡献
+ Cadence	-	-	-	+	-	-	终止式贡献
+ Repeat	-	-	-	-	+	-	长程复现贡献
+ Boundary	-	-	-	-	-	+	局部凝聚贡献
Full - Repeat	+	+	+	+	-	+	检验 A/A’ 关系
Full - Boundary	+	+	+	+	+	-	检验跨边界泄漏
Full MGSA	+	+	+	+	+	+	完整模型

建议额外加入两种**负对照**：

- Random Bias：保持 bias 数值分布但打乱位置；
- Shuffled Prior Labels：检验模型是否只是从附加参数量获益。

如果 Full MGSA 优于 Vanilla，而 Random Bias 无提升，论文说服力会明显增强。

10 论文创新边界与相关工作

当前文献已经覆盖了以下方向，因此不要把这些概念单独宣称为原创：

工作	已有内容	与本项目的边界
Szelogowski et al., ISMIR 2022	Classical form recognition; 200 首作品; form/phrase benchmark	任务领域非常接近, 但并非六类 musicological priors 注入 attention logits。
Peeters, ISMIR 2023	SSM、novelty loss、self-attention 用于 MSA boundary detection	主要是 audio MSA; SSM + self-attention + boundary 已经不是新概念。
Chen & Yoshii, ISMIR 2024	多尺度 temporal self-attention 与 multi-faceted self-similarity	仍主要从音频结构学习 self-similarity, 不是符号古典乐谱的显式音乐学 bias。
Buisson et al., ISMIR 2024	Pairwise link prediction + Graph Attention Network; boundary 与 section labels	说明 pairwise relation + attention 对结构分析有效; 本项目要强调 symbolic priors 和 attention-logit injection。
S3, ISMIR 2024 LBD	Symbolic symphony; form、cadence、harmony、phrase、texture annotations	是本项目的重要数据资源, 不是模型创新。
MU-DiT, IEEE TMM 2026	Form-Guided Attention; motif 到 thematic repetition 的多尺度结构关系	名称和理念接近, 但目标是 symbolic full-song generation; 本项目必须避免泛称“首个 Form-Aware Attention”。

最稳妥的 novelty claim

本项目的潜在新意应放在“**显式音乐学 pairwise priors 直接注入 symbolic-score attention logits, 并同时建模结构吸引与边界抑制, 用于层级古典曲式分析**”这一具体组合上。最终是否构成论文层面的新贡献, 必须由更系统的文献检索与实验结果共同确认。

11 论文结构：程序跑通后再写

推荐论文结构:

1. **Introduction**: 古典曲式分析难点、层级结构、音乐学先验为何重要;
2. **Related Work**: classical form recognition、MSA/SSM、symbolic transformers、structure-aware attention;
3. **Method**: input representation、prior construction、MGSA、multi-task heads、decoder;

4. **Datasets**：数据来源、annotation ontology、split、preprocessing；
5. **Experiments**：baseline、training setup、metrics、ablation；
6. **Results**：主表 + 消融 + 可视化；
7. **Analysis**：A/A′、cadence boundary、长程主题复现 case study；
8. **Limitations**：数据集规模、annotation subjectivity、style dependence、OMR error；
9. **Conclusion**。

11.1 论文主结果表预留模板

Model	Boundary F1	Relation Macro-F1	Form Macro-F1	Params
BiLSTM	-	-	-	-
Vanilla Transformer	-	-	-	-
Transformer + Raw Features	-	-	-	-
MGSA (ours)	-	-	-	-

现在先不要填任何“92%”“SOTA”等数字，除非来自冻结测试集和可复现实验。

12 推荐代码仓库结构

```
xunpu-form-analysis/
  configs/
    baseline.yaml
    mgsa_full.yaml
  data/
    raw/
    processed/
  src/
    parsing/
      musicxml_parser.py
      feature_extractor.py
    priors/
      metric.py
      motif.py
      harmony.py
      cadence.py
      repetition.py
      boundary.py
```

```
  models/
    baseline_transformer.py
    mgsa_attention.py
    hierarchical_encoder.py
    heads.py
  training/
    dataset.py
    losses.py
    trainer.py
    metrics.py
  decoding/
    form_decoder.py
  visualization/
    attention_map.py
    self_similarity.py
tests/
scripts/
  preprocess.py
  train.py
  evaluate.py
  ablate.py
README.md
```

建议所有实验都由配置文件驱动，并保存：Git commit、random seed、dataset split、config、checkpoint、metrics.json。这样后续写论文时不用重新追溯“某个数字是怎么跑出来的”。

13 近期执行清单

P0：现在就做

1. 把当前 prototype 拆成 `parsing / priors / models / training / decoding`；
2. 完成 MusicXML → measure-level cached tensor；
3. 先跑 Vanilla Transformer boundary baseline；
4. 实现 metric、cadence、repetition 三个最有解释力的 prior；
5. 把 MGSA 接进同一训练框架，确保可一键开关 prior；
6. 保存 attention map、SSM 和 boundary probability，可视化一个完整乐章。

P1：baseline 稳定后

1. 加 motif 与 harmony prior；

2. 实现 Same / Variation / Contrast relation pairs；
3. 跑完整 ablation；
4. 做至少 3 个 case study：清楚的 ABA′、奏鸣曲式片段、边界模糊案例；
5. 固定实验表格后再写论文 Method 与 Experiments。

P2：投稿前

1. 文献检索扩展到 symbolic structure analysis、inductive bias attention、music theory informed ML；
2. 检查数据授权、代码许可和可复现性；
3. 增加误差分析与 limitations；
4. 再决定投稿 ISMIR、ICASSP/ISMIR LBD、TASLP/TMM 或其他目标。

研发原则

第一阶段只追求一件事：**得到一个可信的 baseline + MGSA full + ablation 表**。只要这三样成立，论文才真正开始。不要先花大量时间写长篇 Introduction，也不要先做漂亮 demo 替代实验。

14 建议的第一版研究题目

Musicology-Guided Structural Attention for Hierarchical Form Analysis of Symbolic Classical Music

中文可暂定为：

基于音乐学先验结构注意力的符号古典音乐层级曲式分析

如果后续实验发现 boundary inhibition 是最显著贡献，可以进一步把标题收敛到“boundary-aware”或“structure-constrained attention”，使论文问题更加尖锐。

15 参考文献与核验入口

1. D. Szelogowski, L. Mukherjee, B. Whitcomb. *A Novel Dataset and Deep Learning Benchmark for Classical Music Form Recognition and Analysis*. ISMIR 2022, pp. 900–907. DOI/record: https://zenodo.org/records/7416690

2. G. Peeters. *Self-Similarity-Based and Novelty-Based Loss for Music Structure Analysis*. ISMIR 2023, pp. 749–756. https://archives.ismir.net/ismir2023/paper/000089.pdf

3. T.-P. Chen, K. Yoshii. *Learning Multifaceted Self-Similarity Over Time and Frequency for Music Structure Analysis*. ISMIR 2024, pp. 189–197. DOI: https://doi.org/10.5281/zenodo.14877309

4. M. Buisson, B. McFee, S. Essid. *Using Pairwise Link Prediction and Graph Attention Networks for Music Structure Analysis*. ISMIR 2024, pp. 207–214. DOI: https://doi.org/10.5281/zenodo.14877311

5. Z.-S. Lin et al. *S3: A Symbolic Music Dataset for Computational Music Analysis of Symphonies*. ISMIR 2024 Late-Breaking/Demo. Repository: https://github.com/iis-mctl/mctl-symphony-datase

6. Q. Lu, Z. Zhang, C. L. P. Chen, T. Zhang. *MU-DiT: An Efficient Diffusion Transformer for Structure-Aware Symbolic Full-Song Generation*. IEEE Transactions on Multimedia, 2026. DOI: 10.1109/TMM.2026.3727177.

Internal research planning document. Claims of novelty remain provisional until systematic literature review and controlled experiments are complete.