UOT-IR: STRUCTURED ROUTING OF HIGH-POLYPHONY SYMBOLIC MUSIC INTO FIXED-BUDGET REPRESENTATIONS

Ziyue Kang[1,2] **Nan Nan**[*1,2,3] **Chenhao Lin**[1,2,3] **Xiaohong Guan**[1,2,3,4]

[1]Frontier Institute of Science and Technology, Xi'an Jiaotong University, Xi'an, China
[2]Interdisciplinary Research Center of Frontier Science and Technology, Xi'an Jiaotong University, Xi'an, China
[3]MOE KLINNS Lab, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China
[4]Center for Intelligent and Networked Systems, Tsinghua University, Beijing, China
[*]Corresponding author: nan.nan@xjtu.edu.cn

arXiv:2608.00576v1 [cs.SD] 1 Aug 2026

ABSTRACT

High-polyphony symbolic music is increasingly used in generation, analysis, and arrangement, yet many downstream tasks require bounded representations with fixed tracks or slots. Converting richly orchestrated scores into compact forms is therefore necessary, but existing approaches relying on heuristic simplification or generic representation-space reduction often fail to preserve structural roles, orchestration compatibility, and playability under strict budgets. To address the issue, this study reformulates the compression problem as a fixed-budget structured routing problem and proposes Unbalanced Optimal Transport for Information Routing (UOT-IR), a training-free framework based on constrained unbalanced optimal transport. UOT-IR combines an orchestration prior, adaptive marginal relaxation, temporal decoding, and playability-aware projection to produce compact and musically coherent bounded representations. This work further studies two practical settings under the same slot budget: template standardization, which maps each input to a predefined bounded template, and adaptive preservation, which retains representative content without assuming an external template. Experiments on the SymphonyNet corpus show that UOT-IR delivers strong overall performance across both settings, including the best Note-F1 in adaptive preservation (0.9120), together with the lowest structural cost (14.7165) and bad structural confusion rate (0.3406) in template standardization. This work establishes a principled paradigm for fixed-budget symbolic music compression, offering a practical path toward compact, structured, and musically coherent symbolic representations.

1. INTRODUCTION

High-polyphony symbolic music is increasingly used in generation, analysis, arrangement, and archival settings. In practice, however, many symbolic music representations and processing frameworks impose fixed budgets on the number of tracks, parts, or target slots [1–4]. Once a richly orchestrated score exceeds such a structural budget, it must first be converted into a bounded form.

This conversion problem is nontrivial because source tracks may play different musical roles, carry unequal salience, and remain subject to orchestration and playability constraints. Under a fixed budget, some material should be preserved, some should be reassigned to compatible target roles, and some should be discarded. Existing approaches mainly fall into two categories: heuristic simplification and representation-space reduction. Heuristic strategies such as pruning, merging, or rule-based selection can reduce complexity, but they often treat compression as removing less important material, making it difficult to preserve multiple concurrent functions such as melody, accompaniment, and bass in a coordinated way [5, 6]. Representation-space reduction methods can also simplify multitrack content, yet they are not explicitly designed to enforce structured source-to-target assignment or playability constraints such as valid pitch range, polyphonic limits, and performance difficulty [7–10]. Related work on automatic instrumentation further shows that multitrack symbolic music often contains overlapping pitch ranges and dense textures, under which simple assignment strategies can produce unstable or less reliable part allocations [11]. As illustrated in Fig. 1, these limitations lead to three practical failure modes: salient lines may be removed, incompatible materials may be merged into the same bounded slot, and the resulting output may violate basic playability or role-consistency requirements. In particular, NotaGen excludes scores with more than 16 staves because of generation complexity [4].

To address these limitations, we reformulate fixed-budget symbolic compression as a structured routing problem. UOT-IR is a training-free framework based on constrained unbalanced optimal transport that integrates a taxonomy-grounded orchestration prior, symbolic statistical descriptors, adaptive marginal relaxation, temporal coherence, and playability-aware projection to produce compact yet musically coherent bounded outputs. We study two settings under the same slot budget: template standardization, which maps each input into a predefined bounded template, and adaptive preservation, which preserves representative content without assuming an external template. Experiments on SymphonyNet [2] show that UOT-IR achieves favorable overall results against heuristic, representation-space, and simplified transport baselines across both settings.

Our contributions are threefold: (1) we formulate fixed-budget symbolic compression as a structured routing problem; (2) we develop a training-free framework that integrates orchestration prior, constrained UOT, adaptive re-

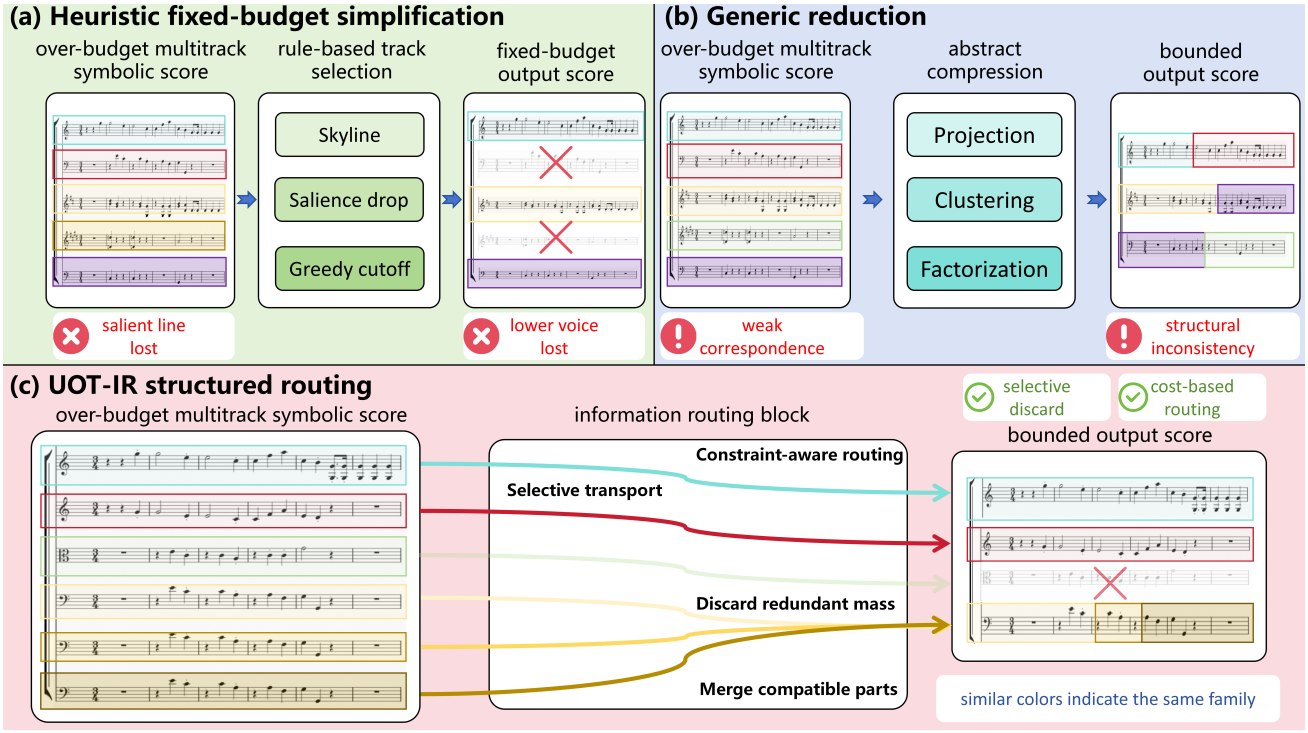

Figure 1. Comparison between conventional fixed-budget compression and UOT-IR. Existing heuristic or representation-space reduction methods often drop salient lines, merge incompatible materials, or violate slot and playability constraints. UOT-IR instead performs structure-aware routing with compatibility-aware assignment, selective discard, and constrained projection to produce coherent bounded outputs.

laxation, temporal decoding, and playability-aware projection; and (3) we evaluate the framework in both settings and show a strong overall balance across fidelity, structural compatibility, and conflict-related metrics.

2. RELATED WORK

2.1 Bounded Symbolic Music Representations

Many symbolic music generation, arrangement, and analysis frameworks rely on bounded structural representations, such as fixed numbers of tracks, parts, or target slots, to simplify storage, batching, alignment, and model design [1–4, 12]. While practical, such representations create a mismatch between richly orchestrated source scores and normalized bounded formats. Existing work typically assumes that the bounded target format is already given, and therefore focuses on modeling, generation, or rearrangement within that format. By contrast, our focus is the conversion problem that arises before those stages: how to transform over-budget multitrack symbolic music into a bounded representation when the source score exceeds the allowed structural budget.

2.2 Heuristic and Representation-Space Reduction

A common way to handle over-budget multitrack music is heuristic reduction, including pruning, merging, or selecting tracks according to activity, density, salience, instrument grouping, or incremental pitch-space coverage [3, 11, 13]. Another line performs representation-space approximation through projection, factorization, or clustering, including principal component analysis (PCA), non-negative matrix factorization (NMF), and K-means [7–9]. These approaches reduce complexity, but they usually do not explicitly model source-to-slot correspondence, orchestration compatibility, and selective discard under a fixed symbolic budget.

2.3 Optimal Transport for Cost-Aware Partial Matching

Optimal transport (OT) studies how to align two distributions under an explicit transportation cost, and has become a general framework for cost-aware soft assignment between structured objects [14, 15]. When strict mass conservation is too restrictive, unbalanced optimal transport relaxes the marginal constraints and allows only part of the source mass to be preserved or reassigned [16, 17]. These properties make UOT a natural tool for symbolic compression under constrained structural budgets, where correspondence, reassignment, and selective discard must all be modeled explicitly.

3. METHOD

UOT-IR instantiates the structured-routing view of fixed-budget symbolic compression. Given an over-budget multitrack symbolic score, the goal is to construct a bounded symbolic output that preserves musically salient content,

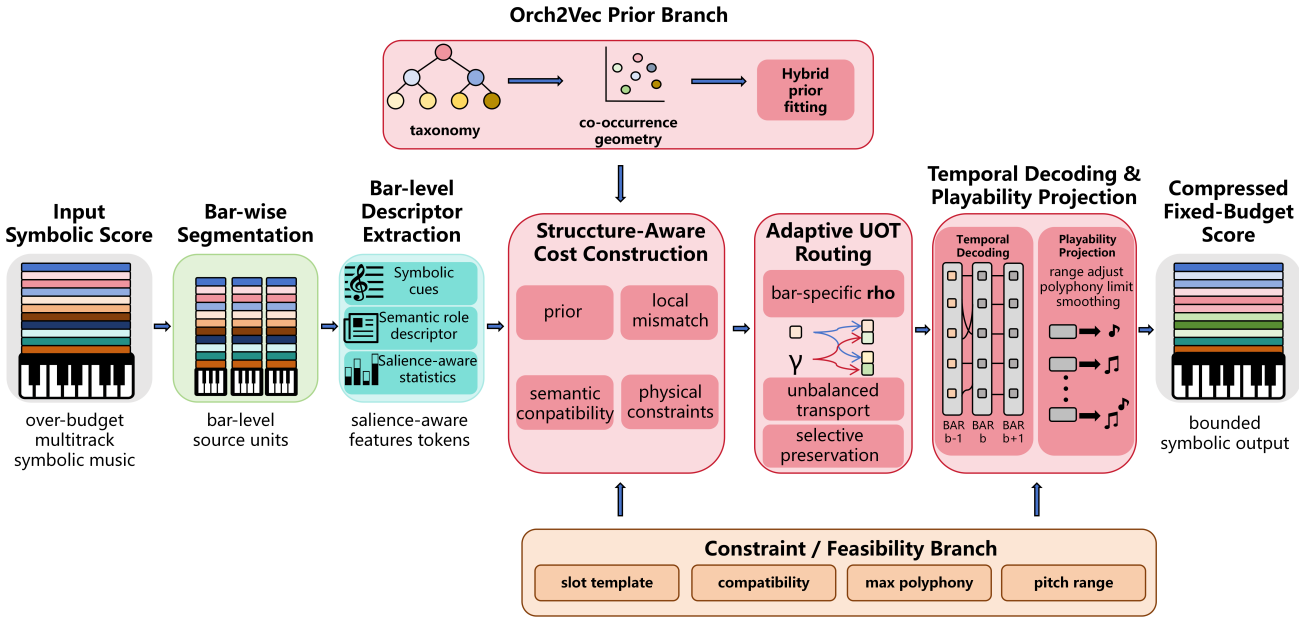

Figure 2. Overview of UOT-IR. The framework consists of five conceptual stages: problem formulation, descriptor and marginal construction, structure-aware cost construction, adaptive UOT routing, and temporal decoding with playability-aware projection. These stages are described across the subsections of Section 3.

maintains structural consistency, and reduces physically implausible assignments, as illustrated in Fig. 1. Instead of treating this task as heuristic track reduction or generic representation-space approximation, the method models it as a structure-aware routing problem under constrained unbalanced optimal transport (UOT). An overview of the full pipeline is shown in Fig. 2.

3.1 Problem Formulation

We segment a multitrack symbolic piece into bars indexed by $b \in \{1, \dots, B\}$, where B is the total number of bars. For bar b, let $X^{(b)} = \{x_1^{(b)}, \dots, x_{N_b}^{(b)}\}$ denote the set of active source tracks and $Y^{(b)} = \{y_1^{(b)}, \dots, y_K^{(b)}\}$ denote the set of K target slots under the fixed compression budget, where K is constant across all bars. In *template standardization*, these slots are predefined by a canonical bounded template. In *adaptive preservation*, they are instantiated from representative source-derived content under the same budget.

For each bar, UOT-IR solves for a nonnegative transport matrix $\Gamma^{(b)} \in \mathbb{R}_+^{N_b \times K}$, where $\Gamma_{ik}^{(b)} \geq 0$ denotes the transported mass from source track $x_i^{(b)}$ to target slot $y_k^{(b)}$. Quantitatively,

$$\begin{aligned} s_i^{(b)} &= \sum_{k=1}^{K} \Gamma_{ik}^{(b)}, \quad c_k^{(b)} = \sum_{i=1}^{N_b} \Gamma_{ik}^{(b)}, \\ m_b &= \|\Gamma^{(b)}\|_1 = \sum_{i=1}^{N_b} \sum_{k=1}^{K} \Gamma_{ik}^{(b)}. \end{aligned} \tag{1}$$

Here, $s_i^{(b)}$ is the retained mass of source track $x_i^{(b)}$, $c_k^{(b)}$ is the received mass of target slot $y_k^{(b)}$, and m_b is the total retained mass at bar b.

3.2 Descriptor and Marginal Construction

Each source track $x_i^{(b)}$ is represented by a bar-level descriptor, while each target slot is associated with a piece-level prototype descriptor

$$\phi_i^{(b)} \in \mathbb{R}^d, \qquad \psi_k \in \mathbb{R}^d, \tag{2}$$

respectively, where d is the shared descriptor dimension. The source descriptors summarize bar-level symbolic statistics, including pitch, duration, onset, density, polyphony, and rhythmic characteristics. To determine how much material should be routed from each source track, we define a nonnegative source marginal

$$\mu^{(b)} \in \mathbb{R}_+^{N_b}. \tag{3}$$

Here, $\mu_i^{(b)}$ is the source mass of track $x_i^{(b)}$, and $\nu_k^{(b)}$ is the receiving mass of target slot $y_k^{(b)}$. In implementation, $\mu^{(b)}$ is computed from bar-level activity and salience cues and then normalized, while the target marginal is uniform, i.e., $\nu_k^{(b)} = 1/K$.

A key component of UOT-IR is **Orch2Vec**, a program-level orchestration prior. We define a vocabulary of 129 instrument tokens, consisting of 128 General MIDI programs and one dedicated drums token. These tokens are organized in a tree-structured taxonomy spanning root-level families (e.g., Strings, Winds, Keys_Plucked, Percussion_Family, and Synth_SFX), intermediate subfamilies, and leaf-level program identities. To complement this symbolic structure, we derive a data-driven program geometry from corpus-level co-occurrence statistics. Let a and

b denote two instrument programs, $P(a,b)$ their joint co-occurrence probability, and $P(a)$ and $P(b)$ their marginal probabilities. We compute the positive pointwise mutual information (PPMI) [18] as

$$\mathrm{PPMI}(a,b) = \max\left\{\log\frac{P(a,b)}{P(a)P(b)},\, 0\right\}. \tag{4}$$

In implementation, the empirical co-occurrence geometry is further stabilized by support gating and clipping before pairwise distances are computed. We then combine this empirical geometry with the taxonomy graph by fitting tree-consistent edge weights and deriving a symmetric prior matrix

$$C_{\text{prior}} \in \mathbb{R}^{129\times 129}, \tag{5}$$

which assigns lower routing cost to source–target program pairs that are both statistically compatible and structurally close in the taxonomy.

3.3 Structure-Aware Cost Construction and UOT Routing

This prior matrix serves as the orchestration-aware component in the subsequent routing-cost decomposition, where it is combined with local, semantic, and physical terms. Given the routing cost matrix

$$C^{(b)} \in \mathbb{R}^{N_b\times K}, \tag{6}$$

we solve for a nonnegative transport matrix by

$$\begin{aligned}\min_{\Gamma^{(b)}\geq 0}\quad & \left\langle \Gamma^{(b)}, C^{(b)}\right\rangle + \lambda_s D_{\rho_b}\left(\Gamma^{(b)}\mathbf{1}, \mu^{(b)}\right)\\ & + \lambda_t D_{\rho_b}\left((\Gamma^{(b)})^\top\mathbf{1}, \nu^{(b)}\right),\end{aligned} \tag{7}$$

where $\langle \Gamma^{(b)}, C^{(b)}\rangle$ is the total routing cost, $\mathbf{1}$ is an all-ones vector, and λ_s and λ_t weight deviations of the transported source and target marginals, $\Gamma^{(b)}\mathbf{1}$ and $(\Gamma^{(b)})^\top\mathbf{1}$, from $\mu^{(b)}$ and $\nu^{(b)}$, respectively. The parameter ρ_b controls marginal relaxation, allowing selective discard under the fixed target budget. A standard choice for the unbalanced penalty is the generalized Kullback–Leibler divergence

$$D_{\rho_b}(\mathbf{a},\mathbf{b}) = \rho_b\sum_j\left(a_j\log\frac{a_j}{b_j} - a_j + b_j\right), \tag{8}$$

where $\mathbf{a}$ and $\mathbf{b}$ are nonnegative vectors of equal dimension, a_j and b_j are their j-th entries, and $\rho_b > 0$ controls the strength of marginal relaxation.

The routing cost matrix is conceptually decomposed as

$$C^{(b)} = \alpha C^{(b)}_{\text{prior}} + \beta C^{(b)}_{\text{local}} + \gamma C^{(b)}_{\text{semantic}} + \delta C^{(b)}_{\text{physical}}. \tag{9}$$

Here, the four components capture program compatibility, descriptor mismatch, role mismatch, and physical infeasibility, respectively, while the conceptual weights $\alpha,\beta,\gamma,\delta \geq 0$ control their relative contributions. In practice, these components are fused through a semantic–local fusion weight, prior–semantic gating, and penalty scaling, which are partially refined by lightweight test-time adaptation (TTA).

The prior term is

$$(C^{(b)}_{\text{prior}})_{ik} = C_{\text{prior}}[p_i, q_k], \tag{10}$$

where p_i and q_k denote the MIDI programs of source track $x_i^{(b)}$ and target slot $y_k^{(b)}$, respectively. The local term is

$$(C^{(b)}_{\text{local}})_{ik} = d_{\text{feat}}\left(\phi_i^{(b)}, \psi_k\right), \tag{11}$$

where d_{feat} is implemented as cosine distance. Thus, Equation (11) measures the distance between the descriptor $\phi_i^{(b)}$ of source track $x_i^{(b)}$ and the prototype descriptor ψ_k of target slot y_k. In the current implementation, target prototypes are initialized from piece-level source-track descriptors and then used as the slot-side references for bar-level routing.

The semantic term is

$$(C^{(b)}_{\text{semantic}})_{ik} = d_{\text{sem}}(r_i^{(b)}, u_k^{(b)}), \tag{12}$$

where $r_i^{(b)}$ and $u_k^{(b)}$ denote semantic embeddings derived from the source-track and target-slot names, respectively, thereby providing coarse role-related cues for routing. The physical term is

$$(C^{(b)}_{\text{physical}})_{ik} = \eta_{\text{range}}\Pi_{\text{range}}(x_i^{(b)}, y_k^{(b)}), \tag{13}$$

where the routing-stage physical term captures pitch-range compatibility between the source material and the target slot.

$$\begin{aligned}\Pi_{\text{range}}(x_i^{(b)}, y_k^{(b)}) = {} & \max(0,\, \ell_k - p^{(b)}_{i,\min})\\ & + \max(0,\, p^{(b)}_{i,\max} - u_k),\end{aligned} \tag{14}$$

where $[p^{(b)}_{i,\min}, p^{(b)}_{i,\max}]$ is the observed pitch interval of the source track and $[\ell_k, u_k]$ is the admissible pitch range of the target slot.

3.4 Adaptive Relaxation, Temporal Decoding, and Playability-Aware Projection

Different bars may require different retention strength, so using a single fixed unbalancedness parameter for all bars is suboptimal. UOT-IR therefore uses a bar-specific relaxation parameter ρ_b. We choose ρ_b by matching the retained mass to a desired target level:

$$\rho_b^\star = \arg\min_{\rho\in\mathcal{R}}\left|\|\Gamma^{(b)}(\rho)\|_1 - \tau_b\right|, \tag{15}$$

where $\rho_b^\star$ is the selected relaxation strength, $\mathcal{R}$ is the search range, $\Gamma^{(b)}(\rho)$ is the transport plan obtained with candidate ρ, and τ_b is the desired retained mass for bar b. In implementation, the current solver performs a lightweight secant-style search in log-ρ space on top of the refined cost matrix, so that each bar obtains its own relaxation strength before the final UOT solve.

Because the transport problem in Equation (7) is solved independently for each bar, the resulting slot identities may fluctuate over time. We therefore convert the bar-wise routing solution into temporally coherent slot sequences

through a target-aware sticky decoding scheme, implemented as a Viterbi-style dynamic program [19]:

$$\max_{\{z_i^{(b)}\}} \sum_{b,i} E_b(i, z_i^{(b)}) - \sum_{b>1,i} T\left(z_i^{(b-1)}, z_i^{(b)}\right), \quad (16)$$

where $z_i^{(b)}$ denotes the target slot assigned to source track i at bar b, and $E_b(i,\cdot)$ is the corresponding emission score derived from the routing result. We define the transition cost between target slots a and b as

$$\begin{aligned} T(a,b) = \lambda_{\text{stay}}\, \mathbf{1}[a \neq b] \\ + \lambda_{\text{prog}}\, \mathbf{1}[\text{prog}(a) \neq \text{prog}(b)]. \end{aligned} \quad (17)$$

It penalizes unnecessary slot switching and target-program changes. During decoding, its strength is further modulated by track importance derived from the accumulated source marginal, improving stability for musically salient tracks.

After temporal decoding, routed note events are projected into target slots with playability-aware refinement, including pitch-range correction, short-note filtering, and target-dependent polyphony control. The final compressed symbolic output for bar b is

$$\hat{Y}^{(b)} = \{\hat{\mathcal{N}}_1^{(b)}, \ldots, \hat{\mathcal{N}}_K^{(b)}\}, \quad (18)$$

where $\hat{\mathcal{N}}_k^{(b)}$ denotes the refined note set assigned to target slot k in bar b. Algorithm 1 summarizes the full UOT-IR pipeline.

Algorithm 1 UOT-IR for fixed-budget symbolic compression

Require: Multitrack symbolic piece $\mathcal{X}$, slot budget K, target regime, orchestration prior C_{prior}
Ensure: Bounded symbolic output $\hat{\mathcal{Y}}$
Segment $\mathcal{X}$ into bars
Instantiate target slots according to the regime
for each bar b **do**
  Extract active source tracks and bar-level descriptors
  Compute source and target marginals
  Build and refine the routing cost matrix
  Search bar-specific relaxation parameter ρ_b
  Solve constrained UOT to obtain $\Gamma^{(b)}$
end for
Apply temporal decoding across bars
Project routed note events with playability-aware refinement
Return $\hat{\mathcal{Y}}$

4. EXPERIMENTAL SETUP

4.1 Dataset and Evaluation Protocol

We evaluate UOT-IR on the SymphonyNet corpus under a fixed slot budget. From the full symbolic collection, we construct evaluation subsets by selecting pieces or bars whose active track count exceeds the target budget, so that compression is required rather than trivial identity mapping. We consider two settings under the same budget K. In **template standardization**, each input is mapped to a predefined bounded target template. In **adaptive preservation**, each method compresses the same input under the same budget without assuming an external template. Unless otherwise stated, all compared methods operate on the same segmented bar-level inputs to ensure fair comparison across settings.

4.2 Baselines

We compare UOT-IR against three groups of baselines. **Heuristic baselines** include *Direct*, *Random*, *Greedy*, *Skyline*, and *BMF-PC*. These are task-level reference strategies constructed for this study, covering direct retention, random selection, salience-driven selection, rule-based reduction, and incremental pitch-space coverage. Specifically, Skyline prioritizes the drum track and the source tracks at the lower and upper register extremes, and then fills the remaining budget according to note density. Boolean matrix pitch-coverage (BMF-PC) is a coverage-based heuristic that iteratively selects representative tracks according to their incremental pitch-space contribution. **Representation-space baselines** include *PCA* [7], *NMF* [8], and *KMeans* [9], which select representative source tracks through projection, factorization, or clustering in a shared structured feature space. These methods are then converted into bounded symbolic outputs for evaluation. **Transport baselines** include *Vanilla-UOT* [16, 17], which removes the structure-aware components from our framework, and *UOT-IR-Core*, which retains the transport backbone but omits selected refinement modules.

In template standardization, methods that do not natively produce slot-aligned outputs are post-aligned to the target template using Hungarian matching [20]. In adaptive preservation, no such post-alignment is applied.

4.3 Metrics

We evaluate all methods from two perspectives. **Content fidelity** is measured by note-level precision, recall, and F1 (Note-P, Note-R, and Note-F1), together with Fréchet Track Embedding Distance (FTED) and Jensen–Shannon distance (JSD) over pitch-class, duration, and inter-onset-interval distributions (PC-JSD, Dur-JSD, and IOI-JSD). **Structural compatibility** is measured by structural cost (SC), bad structural confusion rate (BC), pitch-range difference (PR-Diff), and pitch-class-entropy difference (PCE-Diff). SC measures the average Orch2Vec program cost between matched source and output notes, while BC measures the rate of matched notes whose Orch2Vec program cost exceeds the structural-confusion threshold. Lower FTED, JSD, SC, BC, PR-Diff, and PCE-Diff values indicate better preservation or structural compatibility.

5. RESULTS AND ANALYSIS

5.1 Results on Adaptive Preservation

Table 1 reports results on adaptive preservation. UOT-IR achieves the best overall Note-F1 and recall, indicating a stronger balance between content retention and bounded-output coherence. Heuristic baselines often obtain very high precision by keeping only low-risk content, but their lower recall limits overall usefulness. Compared with representation-space baselines, the gains

suggest that fixed-budget symbolic compression requires correspondence-aware routing rather than generic low-dimensional approximation alone.

Table 1. Main results on the adaptive reduction task.

	Content Fidelity			Distributional and Structural Statistics			
Method	Note-F1* ↑	Note-P ↑	Note-R† ↑	FTED ↓	PC-JSD ↓	Dur-JSD ↓	IOI-JSD ↓
Direct	0.7709	0.9999	0.6563	0.0184	0.1049	0.1282	0.1328
Random	0.7757	0.9999	0.6545	0.0099	0.0768	0.0790	0.1094
Greedy	0.9113	0.9998	0.8458	0.0207	0.0376	0.0562	0.0411
Skyline	0.8778	0.9975	0.7982	0.0138	0.0369	0.0618	0.0481
PCA	0.7092	0.9998	0.5813	0.0071	0.0787	0.0918	0.1244
KMeans	0.7338	0.9999	0.6046	**0.0054**	0.0645	0.0711	0.1139
BMF-PC	0.8989	0.9998	0.8270	0.0172	0.0356	0.0554	0.0444
NMF	0.7401	0.9999	0.6152	0.0060	0.0672	0.0782	0.1102
Vanilla-UOT	0.8593	0.8896	0.8419	0.0351	0.0109	0.0234	0.0130
UOT-IR-Core	0.9053	0.9104	0.9082	0.0228	**0.0041**	**0.0142**	**0.0056**
UOT-IR	**0.9120**	0.9129	**0.9200**	0.0230	0.0071	0.0178	0.0078

Several observations are worth highlighting. First, some baselines achieve nearly perfect precision because they retain only highly confident or low-risk content, but this comes at the cost of much lower recall and therefore lower overall usefulness under fixed-budget preservation. Second, the transport-based variants achieve low values on several distributional metrics, showing that transport-based routing effectively preserves coarse bar-level statistics. However, their weaker overall balance indicates that matching distributions is not sufficient: compression also requires musically structured routing and robust content retention. Third, the full UOT-IR model achieves the best overall Note-F1 and recall, showing that the complete routing formulation provides a strong trade-off between preserving important material and maintaining coherent bounded outputs.

5.2 Main Results on Template Standardization

Table 2 reports results on template standardization. UOT-IR achieves the best overall task-relevant performance, especially on PR-Diff, PCE-Diff, SC, and BC. Although it does not always obtain the lowest FTED, this is expected because template standardization requires not only structural similarity but also template compatibility and low conflict. The strong gap over Vanilla-UOT further shows that transport alone is insufficient under strict template constraints.

Table 2. Main results on the standardization task.

	Fidelity	Standardization and Structural Compatibility				
Method	Note-F1 ↑	PR-Diff* ↓	PCE-Diff* ↓	FTED ↓	SC* ↓	BC* ↓
Direct	0.6712	0.1421	0.0699	0.0186	20.3849	0.4681
Random	0.6732	0.0953	0.0824	0.0100	18.1475	0.4247
Greedy	0.7585	0.0514	0.0745	0.0209	19.1898	0.4645
Skyline	0.7301	0.0568	0.0864	0.0140	18.2123	0.4451
PCA	0.6078	0.1218	0.1210	0.0073	16.6987	0.4116
KMeans	0.6327	0.1009	0.1041	**0.0056**	16.5824	0.3939
BMF-PC	0.7515	0.0501	0.0799	0.0173	18.8827	0.4570
NMF	0.6334	0.1099	0.1077	0.0062	16.8215	0.4077
Vanilla-UOT	0.6335	0.0461	0.0842	0.0294	30.2076	0.7695
UOT-IR-Core	0.9334	0.0196	0.0400	0.0225	18.1452	0.4466
UOT-IR	**0.9370**	**0.0116**	**0.0205**	0.0238	**14.7165**	**0.3406**

A notable result is that UOT-IR does not always achieve the lowest FTED. This is reasonable because FTED captures only one aspect of structural similarity, whereas the standardization task also requires template compatibility, low conflict, and musically plausible slot assignment. In this setting, SC and BC are particularly informative, since they quantify the program-level compatibility of matched source and output notes under the bounded target template. The comparison with Vanilla-UOT is also instructive: without the structure-aware components, a plain transport formulation remains insufficient under a strict template constraint, leading to much worse SC and BC despite using the same transport backbone.

Table 3. Ablation results on standardization and adaptive preservation.

	Standardization			Adaptive			
Variant	Note-F1* ↑	SC* ↓	BC* ↓	Note-F1* ↑	FTED ↓	SC* ↓	BC* ↓
UOT-IR	**0.9370**	14.7165	0.3406	0.9120	0.0230	**2.0296**	**0.0575**
w/o Prior	0.8317	26.9149	0.7528	**0.9310**	0.0162	13.4090	0.2643
w/o Symb. Stat.	0.8853	18.6693	**0.3403**	0.8131	0.0257	3.8790	0.0783
w/o UOT	0.9146	18.5173	0.4754	0.7929	0.0237	9.1925	0.3548
w/o Ada. ρ	0.8601	**13.1580**	0.3808	0.8838	**0.0150**	2.8754	0.0614
w/o Temp.	0.9342	14.6388	0.3492	0.8750	0.0189	2.4669	0.0782
w/o TTA	0.9337	18.1408	0.4462	0.7978	0.0236	6.9246	0.2594

5.3 Ablation Study

To verify the contributions of the main components, we conduct ablation studies under both settings. As shown in Table 3, the prior improves structural compatibility, while UOT and TTA are important for adaptive preservation; overall, the full model achieves the best balance across metrics.

6. CONCLUSION

This paper shows that over-budget symbolic music is better modeled as a structured routing problem than as heuristic simplification or generic representation-space reduction. UOT-IR is a training-free framework based on constrained unbalanced optimal transport. It unifies template standardization and adaptive preservation within a single formulation. Experiments on the SymphonyNet corpus show competitive overall results against heuristic, representation-space, and simplified transport baselines in both settings. These results establish a practical routing-based paradigm for converting high-polyphony symbolic music into bounded yet musically meaningful representations. Future work will study bounded routing-based representations for controllable arrangement, orchestration-aware generation, and fixed-budget symbolic music modeling.

7. ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 62303373, T2341003, and 62376210, and in part by the XJTU Research Fund for AI Science under Grant 2025YXYC011.

8. ETHICS STATEMENT

This work uses a third-party symbolic music dataset in accordance with its applicable licenses and usage conditions.

The proposed method produces fixed-budget symbolic music representations for research on music information routing. The interpretation and use of the resulting representations should consider the musical context, dataset characteristics, and target representation settings.

9. AI USAGE STATEMENT

Generative AI tools were used for language polishing, formatting assistance, and routine coding support.

10. REFERENCES

[1] 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*. Online: Association for Computational Linguistics, 2021, pp. 791–800. [Online]. Available: https://aclanthology.org/2021.findings-acl.70/

[2] J. Liu, Y. Dong, Z. Cheng, X. Zhang, X. Li, F. Yu, and M. Sun, "Symphony generation with permutation invariant language model," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference*, 2022, pp. 551–558. [Online]. Available: https://archives.ismir.net/ismir2022/paper/000066.pdf

[3] L. Ou, J. Zhao, Z. Wang, G. Xia, and Y. Wang, "Unlocking potential in pre-trained music language models for versatile multi-track music arrangement," *arXiv preprint arXiv:2408.15176*, 2024. [Online]. Available: https://arxiv.org/abs/2408.15176

[4] Y. Wang, S. Wu, J. Hu, X. Du, Y. Peng, Y. Huang, S. Fan, X. Li, F. Yu, and M. Sun, "NotaGen: Advancing musicality in symbolic music generation with large language model training paradigms," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025, pp. 10 207–10 215. [Online]. Available: https://doi.org/10.24963/ijcai.2025/1134

[5] S.-C. Chiu, M.-K. Shan, and J.-L. Huang, "Automatic system for the arrangement of piano reductions," in *Proceedings of the 2009 11th IEEE International Symposium on Multimedia*, 2009, pp. 459–464. [Online]. Available: https://doi.org/10.1109/ISM.2009.105

[6] E. Nakamura and S. Sagayama, "Automatic piano reduction from ensemble scores based on merged-output hidden markov model," in *Proceedings of the International Computer Music Conference*, 2015, pp. 298–305.

[7] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[8] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999. [Online]. Available: https://doi.org/10.1038/44565

[9] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.

[10] E. Nakamura and K. Yoshii, "Statistical piano reduction controlling performance difficulty," *APSIPA Transactions on Signal and Information Processing*, vol. 7, p. e13, 2018.

[11] H.-W. Dong, C. Donahue, T. Berg-Kirkpatrick, and J. McAuley, "Towards automatic instrumentation by learning to separate parts in symbolic multitrack music," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, 2021, pp. 159–166. [Online]. Available: https://archives.ismir.net/ismir2021/paper/000019.pdf

[12] S. Ji, X. Yang, and J. Luo, "A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–39, 2023.

[13] J. Zhao, G. Xia, and Y. Wang, "Q&A: Query-based representation learning for multi-track symbolic music re-arrangement," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 5878–5886. [Online]. Available: https://www.ijcai.org/proceedings/2023/652

[14] G. Peyré and M. Cuturi, "Computational optimal transport," *Foundations and Trends in Machine Learning*, vol. 11, no. 5–6, pp. 355–607, 2019. [Online]. Available: https://doi.org/10.1561/2200000073

[15] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems*, vol. 26, 2013, pp. 2292–2300.

[16] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, "Unbalanced optimal transport: Dynamic and kantorovich formulations," *Journal of Functional Analysis*, vol. 274, no. 11, pp. 3090–3123, 2018.

[17] ——, "Scaling algorithms for unbalanced transport problems," *Mathematics of Computation*, vol. 87, no. 314, pp. 2563–2609, 2018. [Online]. Available: https://www.ams.org/mcom/2018-87-314/S0025-5718-2018-03303-8/

[18] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, 1990. [Online]. Available: https://aclanthology.org/J90-1003/

[19] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[20] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1–2, pp. 83–97, 1955.