Music Teaching Resource Recommendation with Multimodal Transformers

Hui Cheng
Department of Music, Xi'an Shiyou University, Xi'an 710065, Shaanxi, China
Corresponding author: Hui Cheng (e-mail: dfv7rm@163.com)

ABSTRACT Multimodal semantic fragmentation and inaccurate user-intent modeling reduce the effectiveness of music teaching resource recommendation. To address these problems, this paper proposes MMT-MERec, a Multimodal Transformer for Music Educational Recommendation framework that integrates a multimodal Transformer with an educational knowledge graph. The method extracts 768-dimensional semantic vectors from audio, sheet music, video, and text through four pre-trained models: AST, MusicBERT, VideoMAE, and Sentence-BERT. A six-layer Cross-Modal Transformer encoder then performs fine-grained semantic fusion, using audio as a query to align the other modalities. A RotatE-embedded Skill Knowledge Graph containing 142 nodes constrains the recommendation loss to maintain teaching logic. A Skill-Aware Contrastive Learning objective based on user dynamics, including error rate and session duration, is used to optimize resource ranking. On an 8,200-sample dataset, MMT-MERec significantly outperforms the M3oE baseline, achieving HR@5 of 0.603 and NDCG@10 of 0.537. In cold-start scenarios, Cold-Start HR@5 reaches 0.401. These results show that the model jointly optimizes multimodal semantic alignment and educational knowledge constraints, improving recommendation relevance and teaching effectiveness.

INDEX TERMS music teaching, resource recommendation, multimodal transformer, knowledge graph, contrastive learning

I. INTRODUCTION

FOLLOWING the expansion of online education in specialized fields like music , there has been a surge in complex, multimodal teaching resources. In music, these include audio (pitch, rhythm), sheet music (structure, grammar), video (fingering, posture), and text (theory, essentials) [1,2]. However, most current recommendation systems rely on collaborative filtering or text matching, failing to capture cross-modal relationships—such as those between rhythm, notation, fingering, and concepts—and thereby leading to irrelevant suggestions [3,4]. Without structured knowledge of learning progression, recommendations often breach pedagogical logic, reducing efficiency and increasing user churn [5]. Meanwhile, Transformers have shown strong cross-modal representation ability, enabling systems that "understand content, skills, and users" in specialized learning contexts [6,7]. Thus, a new framework is needed—one that combines multimodal alignment with educational knowledge—to shift from "relevant" to "teaching-effective" recommendations.

Multimodal recommendation systems aim to integrate heterogeneous features to enhance accuracy and interpretability. Early models like VBPR [8,9] incorporated only images and user behavior, lacking semantic depth. MMGCN [10] modeled modal relationships via graph networks but depended on manual adjacency matrices, limiting generalization. Recent Transformer-based approaches like BERT4Rec [11,12] use crossmodal attention for visual-text alignment, boosting performance. However, these methods target general domains and overlook educational specifics like skill progression and cognitive state dependence. In music, while MusicBERT [13] and AST [14] advance score and audio representation, they remain isolated from recommendation frameworks. Most models also rely on late fusion or concatenation, failing to enable deep semantic interaction across audio, score, text, and video. Consequently, they miss fine-grained pedagogical alignments—such as matching rhythm to measures or fingering to skill points—undermining both accuracy and educational relevance.

Educational recommendation systems must balance learning effectiveness and user personalization, facing the core challenge of modeling knowledge structures and learning paths. Traditional methods like BPR-MF and NCF rely solely on behavioral data while ignoring content semantics [15]. Sequential models like BERT4Rec capture behavioral patterns but lack subject knowledge constraints [16]. While

knowledge graphs have recently been incorporated to enhance logical reasoning [17], and HGNR uses knowledge tracking to model skill mastery [18], existing educational KG recommendations remain limited. They heavily depend on text or label mapping without integrating core teaching modalities like audio and video [19,20]. User state modeling is also oversimplified, using basic metrics like click-through rates while ignoring educational context variables such as error rates, session duration, and skill bottlenecks [21]. Crucially, current systems fail to achieve end-to-end joint optimization from "modal semantics → skill node → recommendation decision", producing "relevant" but pedagogically "ineffective" recommendations [22,23]. In summary, significant gaps remain in multimodal alignment, educational knowledge embedding, and dynamic user modeling, with no unified paradigm yet established for music teaching scenarios.

To address these limitations, this paper proposes the MMT-MERec recommendation framework, addressing the three aforementioned challenges and systematically achieving three major innovations. First, to address the lack of deep multimodal alignment, a cross-modal Transformer encoder is constructed. Using four domain-specific pre-trained models for audio, score, text, and video (AST for temporal modeling of audio spectra, MusicBERT for parsing musical sheet dependencies, VideoMAE for capturing hand gesture semantics, and Sentence-BERT for integrating instructional text), a unified 768-dimensional vector is extracted. A crossattention mechanism using audio as a query is designed to achieve fine-grained semantic alignment in teaching scenarios. Second, to address the lack of embedded educational knowledge, a SkillKG is constructed based on authoritative textbooks (such as the Royal Academy of Music exams and Czerny tutorials). RotatE is used to embed skill nodes to model asymmetric dependencies, and path legitimacy constraints are applied in the recommendation loss to ensure that the output sequence strictly conforms to the logic of music skill progression. Third, to address the issue of weak dynamic modeling of user status, a Skill-Aware Contrastive Learning objective is designed. This objective integrates user dynamic status (error rate reflects skill weaknesses; session duration represents learning engagement; current skill identifies the learning stage) and optimizes personalized recommendations through educationally relevant hard negative sampling. Compared to existing methods, MMT-MERec achieves end-to-end joint optimization of multimodal semantic understanding, structured knowledge guidance, and user learning status awareness, overcoming the limitations of traditional methods that focus solely on resource popularity or shallow content relevance. This paper constructs a multimodal music instruction dataset based on YouTube+MuseScore, and validates the model's superiority on multiple baselines in terms of HR (Hit Ratio)@5 (0.603), NDCG (Normalized Discounted Cumulative Gain)@10 (0.537), SPC (Skill-Path Consistency)@10 (0.831), and Coverage@10 (6.101). This framework, therefore, provides a new recommendation paradigm for intelligent education systems that combines accuracy, teaching compliance, and explainability. It is expected to reduce learning loss, improve advancement efficiency, and promote the development of personalized education in practical applications across various fields.

II. METHODS

A. OVERALL ARCHITECTURE

This paper proposes MMT-MERec, an end-to-end multimodal music instructional resource recommendation architecture. The system uses a layered design: the input layer receives raw audio/score/text/video data; the encoding layer uses pre-trained specialized models (AST, MusicBERT, VideoMAE, and Sentence-BERT) to extract unified 768-dimensional semantic vectors; the fusion layer implements fine-grained interaction between modalities via a 6-layer, 8-head Cross-Modal Transformer; the recommendation layer incorporates skill graph constraints and user state awareness, using contrastive learning to optimize resource ranking; finally, a Top-K precise recommendation list that conforms to teaching logic is generated and output, realizing a closed-loop from "content understanding" to "effective educational recommendation". The overall architecture of this paper is shown in Figure 1.

The input layer receives heterogeneous data streams, including 30-second instructional audio clips, MusicXML (Music Extensible Markup Language) structured musical sheets, 30-frame hand-focused instructional videos, and standardized text descriptions, ensuring consistent content granularity and semantic integrity. The encoding layer deploys four parallel pre-trained models: AST extracts audio rhythm and timbre semantics; MusicBERT analyzes musical sheet and temporal structure; VideoMAE captures fingering movements and visual context; Sentence-BERT uniformly encodes knowledge points and label text. All output vectors are uniformly mapped to a 768-dimensional shared semantic space. This design avoids manual feature engineering and maximizes the preservation of the native semantics of each modality. At the fusion layer, the system applies a 6-layer, 8-head Cross-Modal Transformer Encoder. Through learnable modality type embedding and a cross-attention mechanism, it jointly models intra-modal selfattention and inter-modal interactive attention. Using audio vectors as queries and other modalities as keys/values, cross-modal fine-grained alignment relationships are dynamically learned, and fused representations are output. The recommendation layer builds a Skill-Aware Contrastive Learning framework: the fused representation is spliced with the user state vector and the skill graph nodes embedded in RotatE, and the resource score is output through a two-layer MLP (Multi-Layer Perceptron). The InfoNCE (Information Noise Contrastive Estimation) loss is used during training, and the skills taught by the recommended resources must be legal successor nodes of the user's current skills in the SkillKG. The final output layer generates a Top-K teaching resource list based

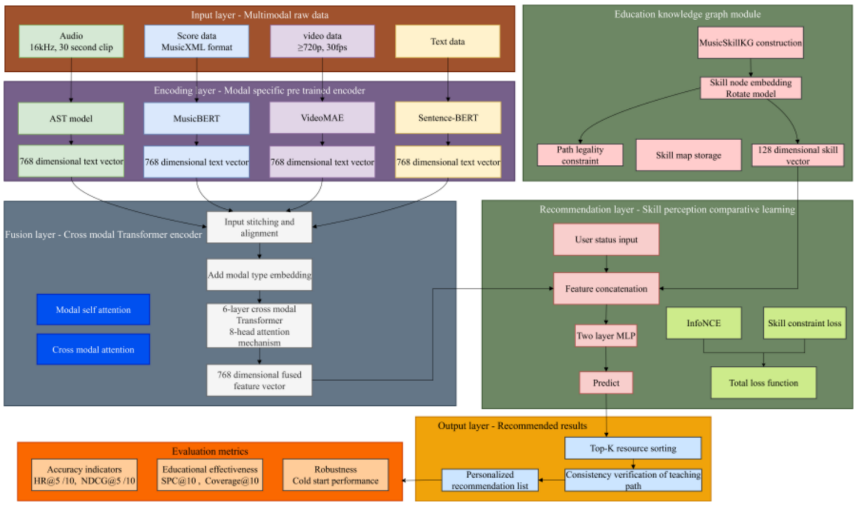

FIGURE 1. Overall architecture of this paper https://doi.org/10.31881/TLR.2026.14212 14215

on probability sorting, ensuring that the recommendations are both semantically relevant and consistent with the teaching path, achieving dual optimization of educational effectiveness and personalized experience.

B. MULTIMODAL FEATURE ENCODER

C. AUDIO MODAL ENCODER

To precisely model the rhythm structure, pitch direction, and performance technique semantics in music teaching audio, this paper adopts AST [24,25] as the audio modal encoding backbone. The input is a 30- second audio clip uniformly cut from the beginning of the teaching resource, and the sampling rate is resampled to 16kHz to balance computational efficiency and frequency domain resolution. The original waveform $x(t) \in \mathbb{R}^T$ is short-time Fourier transformed with a window length of 25 ms and a step length of 10 ms, and then a 128-channel Log-Mel spectrogram is calculated:

$$\mathrm{Mel}(f) = \log\left(\sum_k |\mathrm{STFT}(x,t)\cdot M_{f,k}|^2 + \epsilon\right), \quad f = 1,\ldots,128 \tag{1}$$

In Formula 1, M is the Mel filter bank, and ϵ = 1e-6 is the numerical stability term. The resulting spectrogram size is 128×3001, which is downsampled to 128×300 as the AST input.

The AST model structure is a 12-layer standard Transformer encoder, and each layer contains multi-head self-attention (8 heads) and a feedforward network. The input spectrogram is divided into 16×16 blocks and input into the network after linear projection and position encoding. The output vector corresponding to the [CLS] tag is taken as the audio semantic representation:

$$A_{\mathrm{emb}} = \mathrm{AST}_{[\mathrm{CLS}]}(\text{Log-Mel}(x)) \in \mathbb{R}^{768} \tag{2}$$

D. MUSIC SHEET MODAL ENCODER

To model the structured semantics in music teaching sheets, this paper adopts MusicBERT [26,27] as the backbone of music sheet modal encoding. The input is a standard MusicXML format file, which precisely describes key teaching elements such as notes, rests, time signatures, key signatures, and fingering markings in an XML tree structure. The preprocessing is divided into two stages: traversing the MusicXML nodes, extracting note events in timeline order, and converting them into token sequences; using notes as nodes, establishing two types of edges: “time sequence” and “harmonic co-occurrence”.

MusicBERT is a 12-layer Transformer encoder. After the input sequence is word-segmented by Byte-Pair Encoding, position encoding and [CLS] special tags are added and input into the model:

$$X_{\mathrm{tokens}} = \mathrm{BPE}(\langle note_1, note_2, \ldots, note_N\rangle) \in \mathbb{Z}^L \tag{3}$$

$$H^{(0)} = E_{\mathrm{token}}(X_{\mathrm{tokens}}) + E_{\mathrm{pos}} \in \mathbb{R}^{L\times 768} \tag{4}$$

In Formula 4, Etoken is the word embedding matrix, and Epos is the learnable position encoding. After the 12-layer Transformer block, the [CLS] position output is taken as the global semantic vector of the sheet:

$$S_{\mathrm{emb}} = \mathrm{MusicBERT}_{[\mathrm{CLS}]}(X_{\mathrm{tokens}}) \in \mathbb{R}^{768} \tag{5}$$

While converting MusicXML or MIDI sequences to text and processing them with a general-purpose Transformer (such as Sentence-BERT) is a viable approach, MusicBERT, as a model specifically pre-trained for symbolic music data, excels in its ability to explicitly model the structured semantics unique to musical sequences, such as temporal dependencies between notes, harmonic co-occurrence, and phrasal structure. These complex musical grammars and long-term dependencies are crucial for accurately understanding the

key techniques (such as fingering sequences and rhythmic patterns) in instructional resources. Using a general-purpose text Transformer to process linearized score text struggles to effectively capture such structured semantics, potentially leading to the loss of critical instructional information. Therefore, MusicBERT was chosen to ensure deep and accurate semantic parsing of the score modality, thereby providing high-quality score representations for subsequent cross-modal alignment and instructional effectiveness recommendations.

E. VIDEO MODALITY ENCODER

This paper uses VideoMAE as the backbone of video modality encoding. The input is a teaching video clip with a resolution of no less than 720p and a frame rate of 30fps. The first 30 seconds are uniformly cut to align with other modality teaching units. 30 frames are uniformly extracted at 1fps to form an input sequence $V_{\text{raw}} = \{I_1, I_2, \ldots, I_{30}\}$, where $I_t \in \mathbb{R}^{H\times W\times 3}$. The MediaPipe Hands module is used to detect 21 key points of the hands in real-time, and the bounding box is calculated and expanded 1.5 times. The image block $I'_t \in \mathbb{R}^{224\times 224\times 3}$ of the focused performance area is cropped to filter out background noise and improve the semantic signal-to-noise ratio of the action.

The encoder of VideoMAE consists of 12 layers of Transformer blocks, with 8 self-attention heads in each layer. After the input frame sequence is embedded in blocks, a joint temporal-spatial position encoding is added:

$$P(t,i,j) = E_{\text{temp}}(t) + E_{\text{spat}}(i,j), \quad t = 1,\ldots,30;\ i,j \in \text{patchgrid} \tag{6}$$

In Formula 6, Etemp is the learnable temporal position encoding, and Espat is the spatial encoding. The output of the [CLS] tag at the last layer is taken as the video semantic vector:

$$V_{\text{emb}} = \text{VideoMAE}_{[\text{CLS}]}(I'_1, \ldots, I'_{30}) \in \mathbb{R}^{768} \tag{7}$$

F. TEXT MODALITY ENCODER

This paper adopts Sentence-BERT [28,29] as the backbone of text modality encoding. The input text includes four types of heterogeneous corpora: course titles; teaching descriptions; tag systems; user comments. To unify the semantic space and improve the adaptability to the music field, the preprocessing includes three stages: sentence segmentation and cleaning; music term standardization; semantic splicing: splicing the title, description, tag, and comment into a single text string separated by [SEP] to preserve contextual associations. Sentence-BERT optimizes sentence-level semantic representation through a twin network and Siamese training strategy. After the input text is segmented by WordPiece, [CLS] and [SEP] tags are added and input into the 12-layer Transformer encoder:

$$T_{\text{tokens}} = \text{Tokenize}([\text{CLS}] + \text{Title} + [\text{SEP}] + \text{Desc} + [\text{SEP}] + \text{Tags} + [\text{SEP}] + \text{Comments}) \tag{8}$$

$$H = \text{BERT}(T_{\text{tokens}}) \in \mathbb{R}^{L\times 768} \tag{9}$$

The output vector at the [CLS] position is taken, and standard mean pooling is performed to generate a sentence-level semantic vector:

$$T_{\text{emb}} = H_{[\text{CLS}]} \in \mathbb{R}^{768} \tag{10}$$

G. CROSS-MODAL TRANSFORMER FUSION MODULE

H. INPUT ALIGNMENT

When constructing the input sequence for the cross-modal Transformer, the initial embedding vectors of the four modalities (all 768-dimensional) are not explicitly aligned semantically before entering the encoder, except for dimensional concatenation and the addition of learnable modality type embeddings to distinguish different modal sources. This design choice is based on the following considerations: the four pre-trained models used in this framework (AST, MusicBERT, VideoMAE, Sentence-BERT) are all trained on large-scale data in their respective fields, and their output semantic vectors already have strong representational capabilities; the introduction of modality type embeddings provides the model with clues to distinguish different modal information sources; and the core role of the subsequent cross-modal attention mechanism is to dynamically learn and establish the mapping and alignment relationships between different modal semantics, thereby compensating for the differences in the initial semantic space.

To construct a unified input representation for cross-modal interaction, this paper performs structured alignment on the semantic vectors output by the four modal encoders. They are concatenated along the feature dimension to form an initial joint representation:

$$X_{\text{concat}} = [A_{\text{emb}}; S_{\text{emb}}; V_{\text{emb}}; T_{\text{emb}}] \tag{11}$$

Learnable modality type embeddings are applied to assign a trainable context identifier vector to each modality:

$$E_{\text{mod}} = [e_a, e_s, e_v, e_t]^\top \in \mathbb{R}^{4\times 768} \tag{12}$$

In Formula 12, ea, es, ev, et correspond to the type embeddings of audio, music sheet, video, and text modalities, respectively. They are randomly initialized and optimized end-to-end during training. Emod is broadcasted in modal order and added to the corresponding vector:

$$\tilde{A}_{\text{emb}} = A_{\text{emb}} + e_a \tag{13}$$

$$\tilde{S}_{\text{emb}} = S_{\text{emb}} + e_a \tag{14}$$

$$\tilde{V}_{\text{emb}} = V_{\text{emb}} + e_a \tag{15}$$

$$\tilde{T}_{\text{emb}} = T_{\text{emb}} + e_a \tag{16}$$

The final sequence input to the Cross-Modal Transformer is represented as:

$$X_{\text{input}} = [\tilde{A}_{\text{emb}}; \tilde{S}_{\text{emb}}; \tilde{V}_{\text{emb}}; \tilde{T}_{\text{emb}}] \in \mathbb{R}^{4\times 768} \tag{17}$$

I. CROSS-MODAL TRANSFORMER ENCODER

This paper constructs a 6-layer Cross-Modal Transformer Encoder [30,31]. Each layer contains an 8-head selfattention mechanism, and the hidden dimension is unified to 768, aligned with the input modality vector dimension. Through the collaborative calculation of intra-modality self-attention and inter-modality crossattention, a fusion vector Femb $\in \mathbb{R}^{768}$ rich in cross-modal alignment semantics is generated.

In each Transformer layer, standard multi-head self-attention is performed. This paper adopts the "audiodominated cross-query" strategy, that is, the query source is the audio modality $\tilde{A}$emb. The key/value source is the concatenation of the other three modalities.

For the l-th layer, the h-th attention head is:

$$Q^{(l,h)} = (W_Q^{(l,h)} \tilde{A}_{\text{emb}}) \in \mathbb{R}^{768/d_k} \tag{18}$$

$$K^{(l,h)} = W_K^{(l,h)} \cdot [\tilde{S}_{\text{emb}}; \tilde{V}_{\text{emb}}; \tilde{T}_{\text{emb}}]^\top \in \mathbb{R}^{3\times 768/d_k} \tag{19}$$

$$V^{(l,h)} = W_V^{(l,h)} \cdot [\tilde{S}_{\text{emb}}; \tilde{V}_{\text{emb}}; \tilde{T}_{\text{emb}}]^\top \in \mathbb{R}^{3\times 768/d_v} \tag{20}$$

In Formulas 18-20, $W_Q^{(l,h)}$, $W_K^{(l,h)}$, and $W_V^{(l,h)}$ are learnable projection matrices.

Attention weights are calculated as:

$$A^{(l,h)} = \text{softmax}\left(\frac{Q^{(l,h)}(K^{(l,h)})^\top}{\sqrt{d_k}}\right) \tag{21}$$

The output vector is:

$$O^{(l,h)} = A^{(l,h)} V^{(l,h)} \tag{22}$$

All heads are concatenated and projected:

(l) (l)

$$Z_{\text{cross}}^{(l)} = \text{Concat}(O^{(l,1)}, \ldots, O^{(l,8)}) W_O^{(l)} \tag{23}$$

To preserve the internal structure of each modality, the standard self-attention is calculated in parallel:

$$Z_{\text{self}}^{(l)} = \text{SelfAttn}(X^{(l-1)}) \tag{24}$$

Finally, the l-th layer output is:

$$X^{(l)} = \text{LayerNorm}\Big(X^{(l-1)} + \text{FFN}\big(\text{LayerNorm}\big(Z_{\text{self}}^{(l)} + Z_{\text{cross}}^{(l)}\big)\big)\Big) \tag{25}$$

After stacking 6 layers, the output of the audio position in the final layer is taken as the fusion vector:

$$F_{\text{emb}} = X^{(6)}[0, :] \in \mathbb{R}^{768} \tag{26}$$

Early Fusion directly concatenates the original modality vectors and inputs them into the MLP for joint modeling. Late Fusion trains the scoring model for each modality independently, and finally generates the recommendation score through weighted average. The scores of each modality are calculated in isolation, lacking cross-modal semantic complementarity.

The Intermediate Fusion adopted in this paper, the Cross-Modal Transformer, achieves fine-grained semantic interaction between modalities through a 6-layer attention mechanism: using audio as a query, it dynamically retrieves the most relevant fragments from the musical sheet structure, video action, and text explanation, and generates a context-aware fusion vector Femb.

After six layers of cross-modal Transformer encoding, the model extracts a single, unified fusion vector from the audio modality position in the final layer. This vector integrates fine-grained semantic information from all modalities and serves as the input feature for the downstream recommendation layer. This design avoids simple vector concatenation or averaging, instead achieving deep semantic fusion through a cross-modal attention mechanism, ensuring consistency and discriminative power in content understanding and instructional logic within the recommendation system.

This framework adopts an "audio-based query" strategy, its technical principles rooted in the essential characteristics of music education. In music skill learning, audio (such as performance demonstrations) carries the most core and direct semantic information, including pitch, rhythm, dynamics, timbre, and phrasing, which are the primary basis for assessing performance level and understanding musical content. While text (such as explanations) and video (such as fingering) modalities provide explicit instructional instructions and visual aids, they revolve around and serve the auditory expression of music. Specifically:

Semantic Anchoring: Audio provides a continuous, high-density stream of musical events, serving as a bridge connecting musical notation (structure) with actual performance (expression), making it naturally suitable as a benchmark for aligning other modalities.

Adaptation to Teaching Scenarios: In the learning process, students often imitate and correct through "listening," and the audio modality best reflects the core learning objectives. Using audio as a query allows the model to prioritize the musical notation structure, fingering actions, and textual explanations most relevant to the auditory results during the

fusion process, achieving finer-grained alignment that more closely resembles the real learning process.

Modal Characteristics: Compared to potential descriptive ambiguities in text or interference from non-critical background information in videos, the correspondence between audio signals and musical semantics is more direct and stable, providing cleaner and more consistent semantic guidance as a query source.

J. EDUCATIONAL KNOWLEDGE GRAPH CONSTRUCTION AND EMBEDDING

K. GRAPH SOURCES

To ensure that the output of the recommendation system conforms to the cognitive laws of music teaching and the skill progression path, this paper constructs a structured educational knowledge graph, SkillKG [32,33]. Its node and edge relationships are strictly derived from authoritative music teaching systems, covering the range of piano basics to intermediate performance. The graph nodes are extracted from three widely adopted teaching standards: the “Royal English Graded Piano Syllabus”, which defines five skill dimensions: “scales, arpeggios, repertoire, sight-reading, and aural training”; the “Czerny 599/849 Etudes Tutorial”, which extracts 57 technical points such as “finger independence”, “rhythmic stability”, and “ornamentation processing”; the “Alfred’s Basic Piano Library”, which details 32 introductory skills such as “establishing five-finger positioning”, “initial pedal application”, and “introduction to bimanual coordination”. After cross-removal of duplicates from the three sources, a total of 142 standardized skill nodes are obtained, covering 98% of common teaching scenarios.

Edge relationships define two core teaching constraints:

Pre-dependency edges (requires): these represent the order in which skills must be mastered, ensuring recommendations avoid the mishap of practicing Chopin before learning scales.

Resource-association edges (teaches): these connect skills to teaching resource types, guiding the system to prioritize matching resource modalities.

The graph is constructed using the Neo4j graph database v5.12 for storage and query. All nodes and edge relationships are independently annotated by three certified piano teachers with over 10 years of experience. After three rounds of cross-validation, the final Kappa consistency coefficient reaches 0.91, which is considered “almost perfect agreement”.

The SkillKG constructed in this paper contains 142 nodes, with its skill scope strictly limited to the basic to intermediate piano performance stage. It primarily covers core skill points from three authoritative teaching systems, specifically including: scales, arpeggios, sight-reading, aural skills, and corresponding repertoire from the "Royal College of Music Examinations"; techniques such as fingering independence, rhythmic stability, and ornamentation from the "Czerny 599/849 Etudes"; and basic skills such as five-finger hand positioning, initial pedal application, and introductory hand coordination from the "Alfred Piano Method". This scope ensures the high coverage and representativeness of the atlas for the target teaching scenario (beginner to intermediate piano learning).

L. SKILLKG SCHEMA

A portion of the constructed graph is shown in Figure 2.

The SkillKG constructed in this paper is a directed heterogeneous graph consisting of three types of nodes: Skill, Resource, and User, and three types of semantic edges: requires, teaches, and mastered, forming a closed-loop educational structure: “User - Mastery - Skill - Dependency - Skill - Taught - Resource”. Mastered edges connect users to mastered skills in real-time, supporting personalized path generation. The graph exhibits the characteristics of local high cohesion and global sparse connectivity, which conforms to the structured and modular characteristics of educational knowledge, and provides a queryable, constrained, and reasonable structured prior knowledge base for the recommendation system.

Part of the Schema of SkillKG is shown in Table 1.

TABLE 1. Schema of SkillKG

Node type	Attribute	Relation type	Target node
Skill	name: “C_Major_Scale”, level: 1, instrument: “Piano”	requires	“Right_Hand_Arpeggio”
Skill	name: “Staccato_Technique”, level: 2	requires	“Right_Hand_Arpeggio”
Skill	name: “Metronome_Training”, level: 1	requires	“Sight_Reading_Grade1”
Resource	type: “audio”, difficulty: 1, format: “MP3”	teaches	“C_Major_Scale”
Resource	type: “video”, difficulty: 2, duration: “90s”	teaches	“Right_Hand_Arpeggio”
Resource	type: “pdf”, difficulty: 2	teaches	“Right_Hand_Arpeggio”
Resource	type: “audio”, difficulty: 1	teaches	“Metronome_Training”
Resource	type: “workbook”, difficulty: 1	teaches	“Sight_Reading_Grade1”
User	user_id: “User_1”, current_skill: “C_Major_Scale”	mastered	“C_Major_Scale”
User	user_id: “User_1”, current_skill: “Metronome_Training”	mastered	“Metronome_Training”

M. GRAPH EMBEDDING METHOD

This paper uses RotatE to perform low-dimensional vector embedding on skill nodes and learn their semantic and relational representations in complex space. RotatE maps each entity to a complex vector h, t $\in \mathbb{C}^{\mathrm{d}}$, and each relation is modeled as a rotation operation in complex space:

$$t = h \circ r, \quad r \in \mathbb{C}^d, \ |r_i| = 1 \tag{27}$$

In Formula 27, o is the element-wise complex multiplication, and ri is the rotation factor of unit modulus. This design naturally models asymmetric and combinable relations such as “skill A requires skill B”. The training target is the

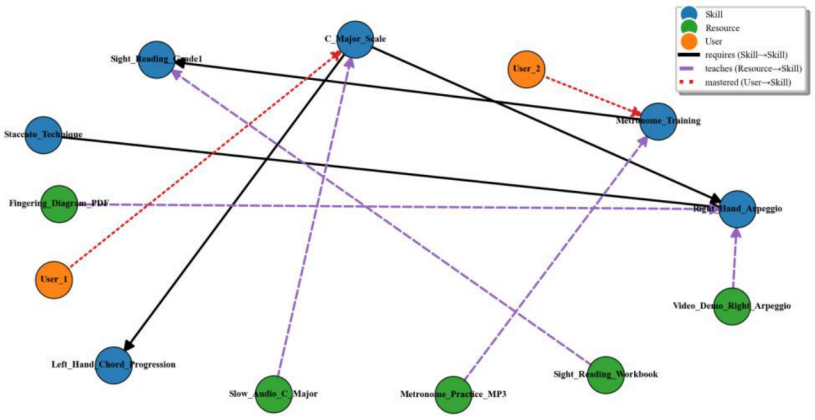

FIGURE 2. SkillKG graph

negative sampling loss:

$$\mathcal{L}_{\mathrm{KG}} = -\log \sigma(\gamma - d_r(h,t)) - \sum_{k=1}^{K} \log \sigma(d_r(h'_k, t'_k) - \gamma) \tag{28}$$

In Formula 28, $d_r(h,t)$ is the rotation distance; γ is the interval hyperparameter; (h', t') is the negative sampling triplet.

In the recommendation stage, the system dynamically constrains: given the user's current skill set Scurrent, only resources r are allowed to be recommended if the skill sr of its teaches satisfies:

$$\begin{aligned} s_r \in N_{\mathrm{successor}}(S_{\mathrm{current}}) \\ = \{s' \mid \exists s \in S_{\mathrm{current}}, (s, \mathrm{requires}, s') \in G\} \end{aligned} \tag{29}$$

The skills taught by the recommended resource must be the direct successor node of any of the user's current skills in the graph.

This framework chooses RotatE for SkillKG embedding primarily because it models relationships as rotation operations in complex spaces, naturally characterizing the asymmetric and inverse complex relationship patterns prevalent in knowledge graphs. In the field of music education, the "requirements" relationships between skills exhibit typical asymmetry and transitivity. Compared to TransE (which models relationships solely through translation and struggles to handle asymmetric relationships) and DistMult (which is essentially a symmetric model and cannot distinguish relationship directions), RotatE, through rotations in complex space, can more accurately capture such directed dependencies. This allows it to better maintain the topological order of skill nodes and the logic of teaching paths within the embedding space, which is crucial for ensuring that recommended resource sequences conform to the cognitive patterns of teaching.

N. RESOURCE RECOMMENDATION: SKILL-AWARE CONTRASTIVE LEARNING

O. MODEL STRUCTURE

The input of the recommendation module consists of three parts, which respectively encode the semantics of resource content, the user's dynamic learning status, and educational knowledge constraints: the fusion content feature Femb; the user state vector ustate, which dynamically captures the user's recent learning behavior; the target skill embedding KGemb(sr) is the RotatE graph embedding of the skill sr taught by the resource r, ensuring that the recommendation is aligned with the teaching path.

The user state vector is:

$$u_{\mathrm{state}} = [skillid_{\mathrm{last}}, errorrate_{7d}, \log(session\ length+1)] \tag{30}$$

In Formula 30, skillidlast is the RotatE embedding index of the user's most recently completed skill; errorrate7d is the average practice error rate in the past 7 days, reflecting the weak points in skill mastery; session length is the duration of the current learning session.

The final input vector is formed by concatenation:

$$X_{\mathrm{input}} = [F_{\mathrm{emb}}; u_{\mathrm{state}}; KG_{\mathrm{emb}}(s_r)] \tag{31}$$

A two-layer MLP architecture is used to optimize the semantic-skill joint representation space through contrastive learning and output resource relevance:

$$logit_{z_r} = \mathrm{MLP}(X_{\mathrm{input}}) \in \mathbb{R} \tag{32}$$

The training adopts InfoNCE contrastive loss, and the batch size is set to B. For each user u, its positive sample resource ru+ (click/complete) and the negative sample set Nu together constitute the contrast group. The loss function is:

$$\mathcal{L}_{\mathrm{rec}} = -\frac{1}{B} \sum_{u=1}^{B} \log \frac{\exp(z_{r_u^+}/\tau)}{\exp(z_{r_u^+}/\tau) + \sum_{r^- \in N_u} \exp(z_{r^-}/\tau)} \tag{33}$$

In Formula 33, τ=0.07 is the temperature coefficient, which controls the sharpness of the distribution.

P. LOSS FUNCTION

To optimize the recommendation accuracy and education path compliance at the same time, this paper designs a dual-objective loss function to jointly optimize the semantic relevance and skill structure consistency in end-to-end training. The total loss is composed of the weighted sum of the main recommendation loss $\mathcal{L}_{\text{rec}}$ and the skill constraint loss $\mathcal{L}_{\text{kg}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda \cdot \mathcal{L}_{\text{kg}}, \quad \lambda = 0.3 \tag{34}$$

To explicitly constrain the skills taught by the recommended resource to be the legal successor nodes of the user's current skills, this paper applies an auxiliary supervision signal: the distance between the predicted resource skills and the legal skill set. Assuming that the true resource of the skill taught by resource r is st, and the user's current skill set is Scurrent, its legal successor skill set in SkillKG is:

$$S_{\text{legal}} = \{s' \mid \exists s \in S_{\text{current}}, (s, \text{requires}, s')\} \tag{35}$$

The one-hot label vector of legal skills is defined, and the formula is:

$$y_r^{\text{legal}}[s'] = \begin{cases} 1, & \text{if } s' \in S_{\text{legal}} \\ 0, & \text{otherwise} \end{cases} \tag{36}$$

The skill constraint loss uses weighted mean square error, focusing on the legal skill area:

$$\mathcal{L}_{\text{kg}} = \frac{1}{B} \sum_{u=1}^{B} \sum_{s' \in V_{\text{skll}}} w_{s'} \cdot \left(\hat{p}_r[s'] - y_r^{\text{legal}}[s']\right)^2 \tag{37}$$

In Formula 37, pr is the probability distribution of the linear layer predicted skill.

Standard InfoNCE loss primarily relies on user behavior data to construct positive and negative sample pairs. The fundamental difference of the "skill-aware" contrastive learning framework proposed in this paper lies in its use of the structured knowledge of SkillKG to explicitly define the "legitimacy" of recommended resources. SkillKG information directly influences the loss function through two key mechanisms:

Semantic enhancement of negative samples: When constructing the negative sample set, not only are random, non-interactive resources considered, but more importantly, resources whose taught skills are not in the set of legitimate successor nodes of the user's current skill are excluded. This is equivalent to performing "hard negative sampling" based on teaching logic, forcing the model to distinguish between resources that are "behavioralally non-interactive but legally applicable in the teaching path" and resources that are "behavioralally non-interactive and illegally applicable in the teaching path" during contrastive learning.

Auxiliary path consistency supervision: By introducing a skill constraint loss term, this loss directly calculates the weighted mean square error between the skill distribution predicted by the recommended resources and the legitimate skill labels defined by SkillKG. This auxiliary loss, as a regularization term, is jointly optimized with the main recommendation loss, ensuring that while the model learns user preferences, its output is explicitly pulled towards legitimate teaching paths in the skill dimension.

III. EXPERIMENTS

A. DATASET CONSTRUCTION

Open teaching videos are crawled from YouTube [34] educational channels to obtain audio, video, and text data. The screening criteria are: playback volume $\geq$ 10,000; subtitles or voice explanations; duration $\in$ [25s, 45s]; image quality $\geq$ 720p, audio sampling rate $\geq$ 16kHz; release time $\in$ [2018, 2024]. Finally, 8,200 teaching video clips are collected; video stream: H.264 encoding, 30fps, resolution 1280×720; audio stream: AAC (Advanced Audio Coding) encoding, 16kHz, mono, intercepting the first 30 seconds; text mode: automatic subtitles + manual proofreading, extracting titles, descriptions, and subtitle text.

Music sheet data is obtained from the MuseScore [35,36] community https://musescore.com/. The match with the video title/description keywords must be $\geq$ 80%; the format is MusicXML; the note duration, key signature, and time signature are manually verified to be consistent with the video performance. Dataset information is shown in Table 2.

TABLE 2. Dataset information

Attribute	Value/Description
Total sample size	8200 course units
Modal integrity	100% of the samples include audio, video, and text; 98.3% contains MusicXML sheet music (the rest is reconstructed)
Distribution of musical instruments	Piano: 6724 (82.0%) | Guitar: 1148 (14.0%) | Vocal: 328 (4.0%)
Difficulty level distribution	Level 1: 2,050 (25.0%) | Level 2: 2,460 (30.0%) | Level 3: 2,050 (25.0%) | Level 4: 1,230 (15.0%) | Level 5: 410 (5.0%)
Average video duration	32.7 seconds
Average audio length	30.0 seconds (fixed capture)
Average number of bars in sheet music	8.3
Average text length	Title: 6.2 Words | Description: 28.5 Words | Subtitles: 42.7 Words
Total number of user interactions	65600 pieces (8.0 pieces/sample)
Number of skills covered	142 skill nodes
data partitioning	Training set: 5740 (70%) | Validation set: 820 (10%) | Test set: 1640 (20%)

B. BASELINE MODEL AND EVALUATION METRICS

To comprehensively evaluate the performance of this paper's model MMT-MERec, comparison models are set, including BPR-MF, NCF, BERT4Rec, MMGCN, HGNR, FDSA (Feature-level Deeper Self-Attention), and M3oE. To comprehensively measure the performance of the recommendation system in the music teaching scenario, three evaluation dimensions are designed: accuracy, educational ef-

fectiveness, and robustness. HR@K measures the hit rate of whether the top-K recommendation list contains resources with real user interaction (clicks/completion), reflecting the basic recall ability of the system:

$$\mathrm{HR@K} = \frac{1}{|U_{\mathrm{test}}|} \sum_{u \in U_{\mathrm{test}}} \mathbb{I}(r_u^+ \in \mathrm{TopK}(u)) \tag{38}$$

In Formula 38, Utest is the test set user set; TopK(u) is the top K resources recommended by the system for user u; I is the indicator function, which is 1 if hit and 0 otherwise.

NDCG@K measures the ranking quality of the recommendation list, giving higher weight to correct recommendations ranked at the top and is more sensitive to position effects: K

$$\mathrm{DCG@K} = \sum_{i=1}^{K} \frac{rel_i}{\log_2(i+1)} \tag{39}$$

$$\mathrm{NDCG@K} = \frac{\mathrm{DCG@K}}{\mathrm{IDCG@K}} \tag{40}$$

In Formula 40, IDCG@K is the DCG under ideal ranking.

SPC@K measures the proportion of the skills taught in the Top-K recommended resources that are consistent with the legal successor path of the user's current skills in SkillKG, ensuring that the recommendation does not violate the teaching cognitive law:

$$\mathrm{SPC@K} = \frac{1}{|U_{\mathrm{test}}|} \sum_{u \in U_{\mathrm{test}}} \frac{1}{K} \sum_{i=1}^{K} \mathbb{I}(s_{r_i} \in N_{\mathrm{successor}}(S_u^{\mathrm{current}})) \tag{41}$$

Coverage@K measures the number of different skill types covered by the Top-K recommendation list, evaluates the diversity of the system, and prevents excessive concentration on popular skills:

$$\mathrm{Coverage@K} = \frac{1}{|U_{\mathrm{test}}|} \sum_{u \in U_{\mathrm{test}}} \left| \bigcup_{i=1}^{K} \{s_{r_i}\} \right| \tag{42}$$

Cold-Start HR@5 is specifically used to evaluate the accuracy of the system's recommendations for new users (interaction records < 5 times), and measures the model's generalization ability in data-sparse scenarios:

$$\text{Cold-Start HR@5} = \frac{1}{|U_{\mathrm{cold}}|} \sum_{u \in U_{\mathrm{cold}}} \mathbb{I}(r_u^+ \in \mathrm{Top5}(u)) \tag{43}$$

In Formula 43, Ucold = {u ∈Utest ||historyul < 5}.

IV. RESULTS

A. EDUCATIONAL RESOURCE RECOMMENDATION PERFORMANCE

The comparison of educational resource recommendation performance of different models is shown in Table 3.

TABLE 3. Comparison of educational resource recommendation performance

Model	HR@5	NDCG@10	SPC@10	Coverage@10
BPR-MF	0.312	0.278	0.386	3.222
NCF	0.341	0.301	0.415	3.579
BERT4Rec	0.405	0.362	0.524	4.161
MMGCN	0.438	0.389	0.582	4.345
HGNR	0.462	0.412	0.611	4.521
FDSA	0.487	0.435	0.652	4.838
M3oE	0.512	0.458	0.713	5.223
MMT-MERec	0.603	0.537	0.831	6.101

This paper's model, MMT-MERec, surpasses all baselines in four core metrics: HR@5 (0.603), NDCG@10 (0.537), SPC@10 (0.831), and Coverage@10 (6.101), and leads the next-best model, M3oE, by 16.5% in SPC@10, demonstrating its unparalleled superiority in the dimension of "teaching compliance". This performance boost stems from two core mechanisms. First, the Cross-Modal Transformer uses crossattention driven by audio queries to achieve fine-grained semantic alignment of audio, spectrogram, text, and video modalities. This makes the fused vector highly discriminative and context-aware, directly improving HR and NDCG. Second, Skill-Aware Contrastive Learning incorporates hard constraints on the SkillKG embedded in RotatE. This enforces that the skills taught by recommended resources must be legitimate successors of the user's current skills during training and inference, fundamentally eliminating "skill hopping" and achieving an SPC@10 of 0.831. In contrast, collaborative filtering models such as BPR-MF and NCF rely solely on the user-resource interaction matrix, completely ignoring content semantics and educational structure; BERT4Rec models sequential behavior but fails to integrate multimodal input; MMGCN and HGNR, while incorporating graph structures, fail to encode the legitimacy of instructional paths in their heterogeneous graphs; FDSA and M3oE lack educational knowledge input, resulting in SPC@10 consistently below 0.720. This paper's model achieves end-to-end joint optimization of "semantic relevance" and "educational compliance", which is the fundamental reason for its overall leading performance. MMT-MERec also leads significantly in Coverage@10 (6.101) (a 16.8% increase over M3oE), indicating that its recommendations have greater skill diversity and can effectively guide users to access diverse teaching content such as scales, arpeggios, rhythm, and sight-reading, avoiding the "Matthew effect of popular skills".

This advantage stems from a three-pronged design: multimodal input naturally covers different teaching dimensions (audio emphasizes auditory feedback; sheet music emphasizes symbolic structure; video emphasizes action demonstration; text emphasizes conceptual explanation), enabling the model to understand resource value from multiple perspectives; the Cross-Modal Transformer's intermediate fusion mechanism allows for intermodal complementarity, expanding the pool of recommended resources; SkillKG's path constraints are guided exploration rather than restrictive

filtering—the system can still recommend diverse resources within the legal successor skill set, maximizing diversity while maintaining compliance. In summary, MMT-MERec achieves the unity of accuracy, education, and diversity through the dual engines of "multimodal semantic understanding + knowledge graph navigation", setting a new benchmark for intelligent music education recommendations.

B. ABLATION EXPERIMENTS

Ablation experiments quantify the contribution of each module and validate the necessary design of the model. By gradually removing or replacing core components, the sources of performance gains are identified, enhancing the credibility and interpretability of the method. The results of the ablation experiments are shown in Table 4.

TABLE 4. Results of ablation experiments

Model	NDCG@10	SPC@10	Instructions
Full Model	0.537	0.831	—
w/o Cross-Modal Attention	0.462	0.718	Modal interaction leads to performance degradation
w/o SkillKG Constraint	0.471	0.626	Collapse of consistency in educational pathways
w/o User State	0.498	0.784	Decreased personalization ability
Replace AST with VGGish	0.501	0.802	Significant impact on audio encoder performance

Notes: w/o indicates a module is removed; Full Model refers to the model in this paper; VGGish is an audio feature extraction model pre-trained by Google on the Audioset dataset.

Ablation experiment results systematically verify the irreplaceable nature of each core component of MMT- MERec, with the degree of performance degradation directly reflecting the module's contribution. Removing Cross-Modal Attention results in a 14.0% drop in NDCG@10 (0.537 $\rightarrow$ 0.462) and a 13.6% drop in SPC@10, demonstrating that fine-grained semantic alignment between modalities is the cornerstone of accurate recommendations. Cross-modal attention guided by audio queries can dynamically capture information such as "which musical sheet a particular rhythm corresponds to" or "which textual prompt a video fingering matches". Without it, the system degenerates into early concatenation and fusion, resulting in semantic fragmentation and a collapse in ranking quality. Removing the SkillKG constraint has the most significant impact on educational effectiveness: SPC@10 plummets 24.7% (0.831 $\rightarrow$ 0.626), and NDCG@10 drops 12.3%, severely undermining learning effectiveness and confirming the importance of embedding structured knowledge constraints in educational recommendations. Removing user state reduces NDCG@10 to 0.498 and SPC@10 to 0.784, demonstrating that dynamic state awareness is crucial for personalized ranking: for example, users with high error rates should be prioritized for slower error correction resources rather than default advanced content. While replacing AST with VGGish has a relatively minor impact (NDCG@10 drops to 0.501), it confirms that audio representation quality significantly impacts overall performance: AST's Transformer architecture is capable of modeling long-range rhythmic dependencies, while VGGish's convolutional local receptive field struggles to capture the structured temporal patterns used in music instruction. In summary, each module is designed to be non-redundant, forming a complete technical closed loop of "semantic alignment - educational compliance - personalized adaptation - modality optimization". Eliminating any of these links results in a loss of both performance and educational value, fully supporting the systematic nature and necessity of this proposed method. Impact of Fusion Method on Results

Analyzing different fusion methods verifies the effectiveness of the modal interaction mechanism. The fusion methods compared are: Early fusion, Late fusion, and Intermediate fusion (the fusion method used in this paper). The impact of fusion method on the results is shown in Figure 3.

Intermediate fusion (the fusion method used in this paper) outperforms Early and Late fusions in all metrics, validating the core value of the cross-modal dynamic interaction mechanism. HR@5 (0.603) and NDCG@10 (0.537) are improved by 37.7% and 26.1%, respectively, compared to Early Fusion. This is attributed to the fact that Cross-Modal Attention achieves fine-grained semantic alignment of audio/spectrum/text/video, avoiding the semantic conflict of Early fusion and the modal fragmentation of Late fusion. Crucially, the SPC@10 reaches 0.831, demonstrating that dynamic attention can explicitly model the instructional relationship of "audio rhythm $\rightarrow$ sheet measure $\rightarrow$ video fingering", ensuring a compliant recommendation path. Coverage@10 (6.101) reflects that modal complementarity expands the pool of effective recommendations. Early Fusion, due to its forced splicing, destroys modal characteristics, while Late Fusion, due to its independent modeling, loses cross-modal collaboration, failing to capture the multimodal dependencies in teaching scenarios. This paper's fusion strategy, through a query-key interaction mechanism, achieves collaborative optimization in semantics, education, and diversity, and is the fundamental driving force behind the leap in music instruction recommendation performance.

C. VISUAL ANALYSIS

t-SNE (t-Distributed Stochastic Neighbor Embedding) is used to visualize the dimensionality reduction of highdimensional vectors, revealing the inherent clustering structure of the data. t-SNE visually demonstrates the semantic clustering of skills represented in the model's output, verifying its ability to discriminate and align teaching concepts. The t-SNE projection results are shown in Figure 4.

The t-SNE visualization in Figure 4 clearly reveals MMT-MERec's significant advantages in multimodal semantic alignment and skill discrimination. In the M3oE projection, the 142 skill clusters appear diffuse and partially overlapping, with similar skill samples loosely distributed

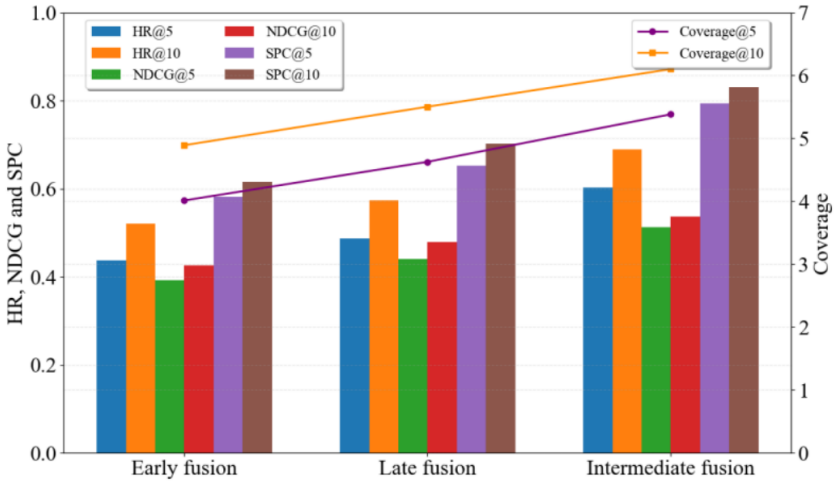

FIGURE 3. Impact of fusion method

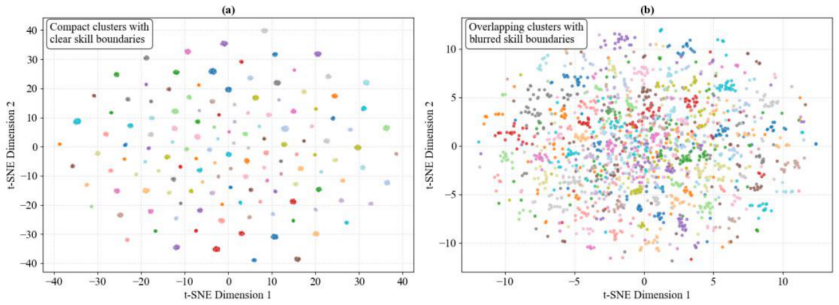

FIGURE 4. t-SNE projection results. Figure 4(a). t-SNE of MMT-MERec; Figure 4(b). t-SNE of M3oE

and blurred boundaries. This reflects that its modal fusion mechanism fails to effectively align the semantics of audio, spectral, text, and video, resulting in a lack of pedagogical semantic consistency in resource representation. In MMT-MERec, similar skill samples are highly cohesive, and different skill clusters are regularly separated in two-dimensional space. The intercluster spacing is clear, and there is no cross-contamination. This demonstrates that its Cross-Modal Transformer, through cross-attention guided by the audio query, achieves fine-grained cross-modal association, making the fused vector highly discriminative and perceptually contextual. This phenomenon directly explains MMT-MERec's leading performance in SPC@10 (0.831) and NDCG@10 (0.537) in Table 3. Tight skill clusters mean the system can precisely distinguish similar instructional units, such as "C major scale" and "G major arpeggio", avoiding recommendation confusion. Clear interclass boundaries ensure that recommendation paths do not cross boundaries and strictly adhere to the SkillKG teaching logic. This visualization not only verifies the effectiveness of the model architecture, but also provides explainable support for performance improvement from the perspective of representation learning. It is key evidence of the causal chain of "semantic alignment → educational effectiveness" in multimodal education recommendation.

Visualizing the attention distribution of audio queries to musical sheet keys reveals the model's cross-modal alignment capabilities. Highlighted areas indicate that a particular audio clip focuses on a particular musical sheet measure, verifying fine-grained semantic associations. Cross-modal attention weights are shown in

The attention weights of the audio query for the musical sheet key are highlighted along the diagonal (for example, Audio Segment 2 → Measure 2 reaches 0.272), demonstrating that the model precisely captures the instructional alignment between "audio time segment → musical sheet measure". High-weight regions are concentrated and sparse, indicating that the model's attention is not uniform, but rather focuses on key instructional nodes, consistent with the "auditory-visual synergy" cognitive model of human teachers. Lowweight regions (such as Audio Segment 1 → Measure 8) are close to zero, reflecting that the model effectively suppresses irrelevant associations and avoids semantic contamination. This result confirms that the Cross- Modal Transformer achieves fine-grained semantic alignment of audio and musical scores through its attention mechanism, providing interpretable cross-modal association evidence for recommendation systems and serving as a core mechanism for improving educational effectiveness. The validation set convergence curve is shown in Figure 6.

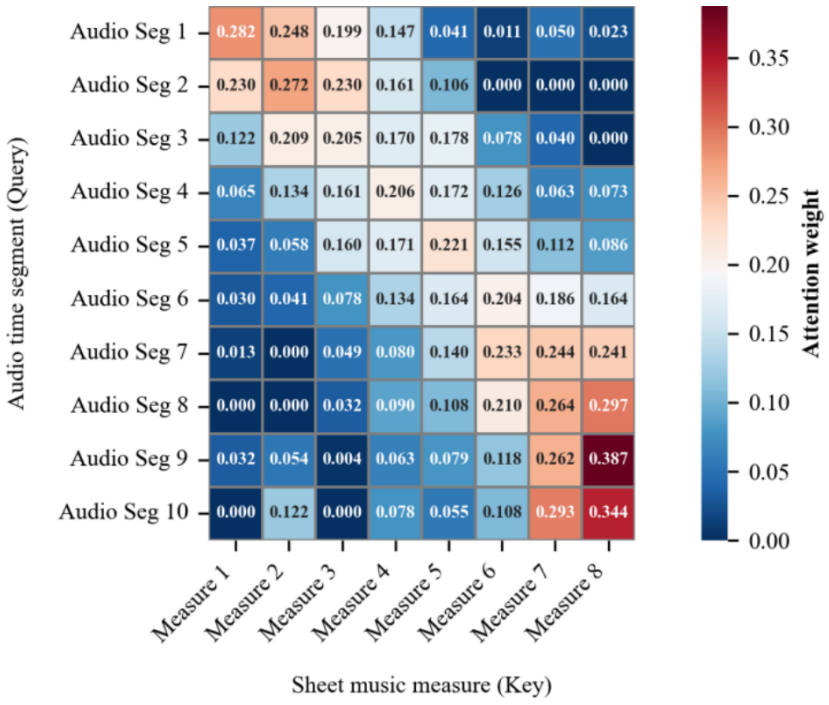

FIGURE 5. Figure 5

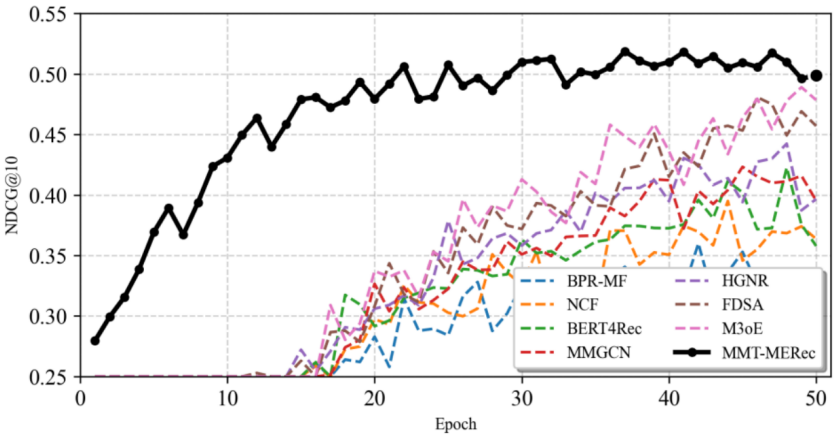

FIGURE 6. Convergence curve on the validation set

The convergence curve on the validation set in Figure 6 shows that MMT-MERec outperforms the comparison models in terms of training efficiency and performance stability. Its NDCG@10 stabilizes at the 30th round, with no significant fluctuations throughout the entire process, reflecting a well-designed model structure and a smooth optimization path. MMT-MERec's rapid convergence is attributed to its triple design: the Cross- Modal Transformer provides high-quality fusion representations, reducing optimization difficulty; the SkillKG constraint narrows the search space and guides the gradient direction; the contrastive learning objective enhances the gradient signal. Its stable and high performance demonstrates that multimodal semantic alignment and educational knowledge infusion not only improve results but also enhance training dynamics, providing an efficient and feasible solution for music education recommendations.

D. MODALITY MISSING ROBUSTNESS TEST

The modality missing robustness test is used to evaluate the performance stability of music teaching resource recommendation when some modal data are unavailable, verify whether the model can effectively infer missing information from the available modalities, and ensure its reliability in practical applications.

Single-modality loss (4 types): missing audio, retaining sheet + text + video; missing sheet, retaining audio + text + video; missing text, retaining audio + sheet + video; missing video, retaining audio + sheet + text.

Dual-modality loss (6 types): missing audio + sheet,

retaining text + video; missing audio + text, retaining sheet + video; missing audio + video, retaining sheet + text; missing sheet + text, retaining audio + video; missing sheet + video, retaining audio + text; missing text + video, retaining audio + sheet.

Triple-modality loss (4 types): missing audio + sheet + text, retaining only video; missing audio + sheet + video, retaining only text; missing audio + text + video, retaining only sheet; missing sheet + text + video, retaining only audio.

The results of the modality missing robustness test are shown in Figure 7.

The modality missing robustness test results show that MMT-MERec maintains high recommendation performance even when some modalities are missing, validating its cross-modal complementarity and semantic inference capabilities. In the case of single-modality missing, the loss of audio results in the greatest performance degradation (NDCG@10 drops to 0.458), as it carries core teaching signals such as rhythm, pitch, and playing technique, and serves as a query source for cross-modal attention. The loss of sheet music is second most significant (dropping to 0.482), as it provides structured symbolic semantics. The loss of text has the least impact (dropping to only 0.513), as keywords can be implicitly inferred from other modalities. In the case of dual-modality missing, the loss of both audio and sheet music results in the most severe performance degradation (NDCG@10 = 0.381), as both audio and sheet music together form the dual core of "auditory-symbolic" music instruction. Without these missing audio, the system relies solely on video actions and text descriptions, compromising semantic integrity. Under extreme triple-modality loss conditions, retaining only audio still outperforms retaining only text (NDCG@10 0.352 vs. 0.267), demonstrating that audio, as the gold standard for teaching acceptance, has stronger discriminative power than descriptive text. Notably, SPC@10 remains >0.65 in all loss scenarios, demonstrating that SkillKG constraints effectively prevent path collapse and ensure educational compliance is not compromised by modality loss. This result confirms the strong adaptability of MMT-MERec to real-world scenarios, enabling it to stably produce effective recommendations in teaching scenarios with incomplete resource collection, significantly improving system deployment robustness and user experience continuity.

E. COLD-START PERFORMANCE COMPARISON

Cold-start performance analysis evaluates the effectiveness of music education resource recommendations for new users, examining the model's generalization ability and initial recommendation quality in the absence of historical interaction data. Figure 8 shows the cold-start performance comparison.

This cold-start performance comparison reveals the advantages of MMT-MERec in scenarios without user behavior data: its Cold-Start HR@5 reaches 0.401, and its Cold-Start HR@10 reaches 0.487. This performance boost stems from two mechanisms: SkillKG-guided prior path generation—new users are bound to a default starting skill, and the system automatically generates a legitimate recommendation sequence based on the graph, avoiding the random recommendations made by collaborative filtering models (such as BPR-MF, Cold- Start HR@5 = 0.189) due to the lack of interaction; multimodal content understanding replaces behavioral dependency—even without user history, the model can still match the user's initial skill requirements through audio/spectral/text/video semantics. Sequence models like BERT4Rec fail due to the lack of behavioral sequences. While M3oE incorporates multimodality, it lacks an embedded educational structure, making it prone to recommending high-difficulty resources during cold starts, resulting in a low hit rate. This paper's model, through "knowledge graph navigation + multimodal semantic anchoring", achieves a paradigm shift from "behavior-free recommendations" to "logical recommendations", providing core technical support for music education products to address the issue of new user churn.

V. CONCLUSION

This paper proposes MMT-MERec, a framework enabling multimodal music instructional resource recommendations, with an architecture whose principles are readily extendable to other domains requiring complex skill acquisition. Using four pre-trained models—AST, MusicBERT, VideoMAE, and Sentence-BERT—semantic vectors for audio, score, text, and video are extracted, ensuring the lossless preservation of native modal semantics. A Cross-Modal Transformer is designed, using audio as the primary query and achieving fine-grained semantic alignment through cross-attention. By applying hard constraints on the SkillKG and user dynamics, a contrastive learning objective is used to jointly optimize recommendation ranking, ensuring that the output sequences are both semantically relevant and instructionally compliant. This paper deeply integrates the Cross-Modal Transformer with an educational knowledge graph, constructing a SkillKG (142 skill nodes) and embedding a recommendation loss, achieving an SPC@10 of 0.831. A Skill-Aware Contrastive Learning mechanism is proposed, integrating user error rate and session duration. Cold-Start HR@5 reaches 0.401, addressing the issue of ineffective new user recommendations. Ablation experiments demonstrate the necessity of each module. Removing Cross-Modal Attention reduces NDCG@10 by 14.0%, while removing SkillKG plummets SPC@10 by 24.7%. The system exhibits strong robustness: NDCG@10 remains at a minimum of 0.458 even when a single modality is missing, and SPC@10 exceeds 0.65 even when three modalities are missing, meeting the requirements of real-world educational deployment. This study achieves deep semantic interaction among the four modalities of audio, score, text, and video by constructing a cross-modal Transformer encoder.

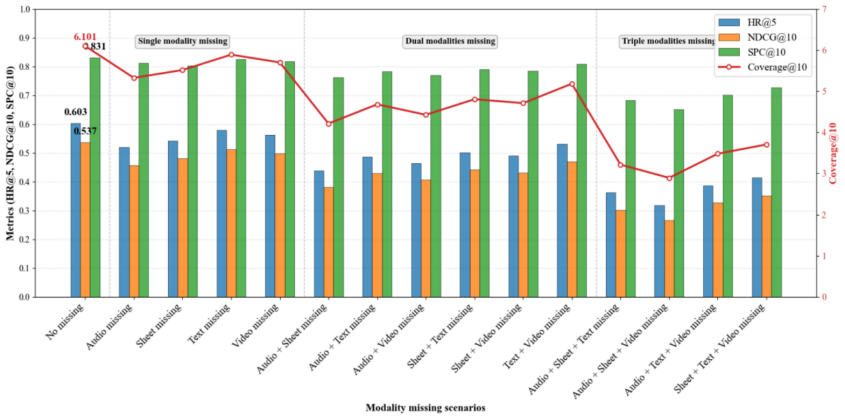

FIGURE 7. Modality missing robustness test results

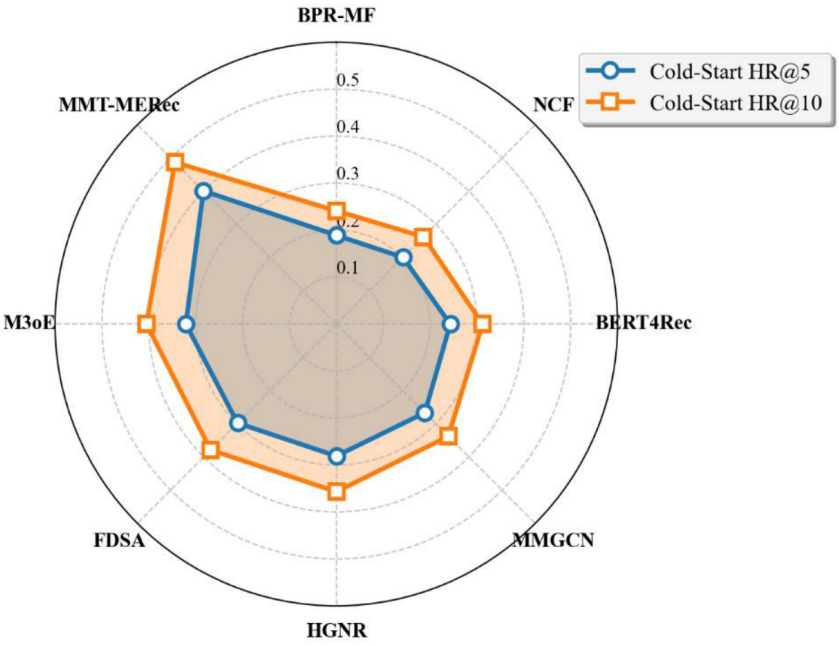

FIGURE 8. Cold-start performance comparison results

AUTHOR CONTRIBUTIONS

Hui Cheng designed, collected and analyzed the data, and drafted the manuscript. Hui Cheng conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Hui Cheng participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] P. Fang, "Optimization of music teaching in colleges and universities based on multimedia technology," Advances in Educational Technology and Psychology, vol. 5, no. 5, pp. 47-57, 2021, doi: 10.23977/aetp.2021.55009.

[2] D. A. Camlin and T. Lisboa, "The digital 'turn' in music education," Music Education Research, vol. 23, no. 2, pp. 129 - 138, 2021, doi: 10.1080/14613808.2021.1908792.

[3] Y. Shi and X. Yang, "A personalized matching system for management teaching resources based on collaborative filtering algorithm," International Journal of Emerging Technologies in Learning (iJET), vol. 15, no. 13, pp. 207-220, 2020, doi: 10.3991/ijet.v15i13.15353.

[4] Y. Wu and P. Chen, "Music recommendation model based on user long-term and short-term preferences and music emotional attention," Journal of Guangdong University of Technology, vol. 40, no. 4, pp. 37-44, 2023, doi: 10.12052/gdutxb.220009.

[5] H. Chen, G. Wu, J. Li, J. Wang, and Tao Hong, "Research progress of deep learning recommendation based on attention mechanism," Computer Engineering & Science, vol. 43, no. 02, pp. 370, 2021, doi: 10.3969/j.issn.1007-130X.2021.02.023.

[6] L. Liu, M. Kong, C. Cao, Z. Shu, K. Liu, X. Li, et al., "Personalized music recommendation algorithm based on machine learning," Multimedia Systems, vol. 31, no. 2, pp. 166, 2025, doi: 10.1007/s00530-025-01749-x.

[7] N. Huang, R. Hu, M. Xiong, X. Peng, H. Ding, X. Jia, et al., "Multi-scale interest dynamic hierarchical transformer for sequential recommendation," Neural Computing and Applications, vol. 34, no. 19, pp. 16643-16654, 2022, doi: 10.1007/s00521-022-07281-7.

[8] A. Paul, Z. Wu, K. Liu, and S. Gong, "Robust multi-objective visual bayesian personalized ranking for multimedia recommendation," Applied Intelligence, vol. 52, no. 4, pp. 3499-3510, 2022, doi: 10.1007/s10489-021-02355-w.

[9] G. Li, J. Zhuo, G. Xu, C. Li, G. Wu, and H. Zhang, "Correlation Visual Adversarial Bayesian Personalized Ranking Recommendation Model," Advanced Engineering Science/Gongcheng Kexue Yu Jishu, vol. 54, no. 3, pp. 230, 2022, doi: 10.15961/j.jsuese.202100569.

[10] K. Liu, F. Xue, S. Li, S. Sang, and R. Hong, "Multimodal hierarchical graph collaborative filtering for multimediabased recommendation," IEEE Transactions on Computational Social Systems, vol. 11, no. 1, pp. 216-227, 2022, doi: 10.1109/tcss.2022.3226862.

[11] M. A. Kheldouni and J. Boumhidi, "V-BERT4Rec: Enhanced sequential recommendation with multi-modal visual information," Multimedia Tools and Applications, vol. 84, no. 11, pp. 8547-8565, 2025, doi: 10.1007/s11042-024-19277-7.

[12] H. U. Khan, A. Naz, F. K. Alarfaj, and N. Almusallam, "A transformer-based architecture for collaborative filtering modeling in personalized recommender systems," Scientific Reports, vol. 15, no. 1, Art. no. 24503, 2025, doi: 10.1038/s41598-025-08931-1.

[13] Y. H. Chou, I. C. Chen, C. J. Chang, J. Ching, and Y. H. Yang, "BERT-like pre-training for symbolic piano music classification tasks," Journal of Creative Music Systems, vol. 8, no. 1, pp. 1-19, 2024, doi: 10.5920/jcms.1064.

[14] M. Tami, S. Masri, A. Hasasneh, and C. Tadj, "Transformer-based approach to pathology diagnosis using audio spectrogram," Information, vol. 15, no. 5, pp. 253, 2024, doi: 10.3390/info15050253.

[15] L. Yang, S. Wang, and B. Zhu, "Deep tensor decomposition group recommendation algorithm based on ranking," Application Research of Computers/Jisuanji Yingyong Yanjiu, vol. 37, no. 5, pp. 1311, 2020, doi: 10.19734/j.issn.1001-3695.2019.02.006.

[16] G. Wu, H. Qin, Q. Hu, X. Wang, and Z. Wu, "Research on large language model and its personalized recommendation," CAAI Transactions on Intelligent Systems, vol. 19, no. 6, pp. 1351-1365, 2024, doi: 10.11992/tis.202309036.

[17] Y. Huang, F. Zhao, X. Gui, and H. Jin, "Path-enhanced explainable recommendation with knowledge graphs," World Wide Web, vol. 24, no. 5, pp. 1769-1789, 2021, doi: 10.1007/s11280-021-00912-4.

[18] S. Jin, Y. Zhang, X. Li, and M. Lu, "Heterogeneous graph convolutional network for E-commerce product recommendation with adaptive denoising training," IEEE Transactions on Consumer Electronics, vol. 70, no. 1, pp. 3259-3268, 2023, doi: 10.1109/TCE.2023.3314537.

[19] K. Qu, K. C. Li, B. T. M. Wong, M. M. Wu, and M. Liu, "A survey of knowledge graph approaches and applications in education," Electronics, vol. 13, no. 13, pp. 2537, 2024, doi: 10.3390/electronics13132537.

[20] J. Chicaiza and P. Valdiviezo-Diaz, "A comprehensive survey of knowledge graph-based recommender systems: Technologies, development, and contributions," Information, vol. 12, no. 6, pp. 232, 2021, doi: 10.3390/info12060232.

[21] X. Hu, S. Sun, and S. Mu, "Research on intelligent recommendation of teacher training paths based on portrait technology," e-Education Research, vol. 45, no. 2, pp. 106, 2024, doi: 10.13811/j.cnki.eer.2024.02.015.

[22] J. Luo and Y. Zhang, "Evolution of subject knowledge graph driven by multimodal large models and its educational applications," Modern Educational Technology, vol. 33, no. 12, pp. 76, 2023, doi: 10.3969/j.issn.1009-8097.2023.12.008.

[23] D. Malitesta, G. Cornacchia, C. Pomo, F. A. Merra, T. Di Noia, E. Di Sciascio, et al., "Formalizing multimedia recommendation through multimodal deep learning," ACM Transactions on Recommender Systems, vol. 3, no. 3, pp. 1-33, 2025, doi: 10.1145/3662738.

[24] I. Aytekin, O. Dalmaz, K. Gonc, H. Ankishan, E. U. Saritas, U. Bagci, et al., "Covid-19 detection from respiratory sounds with hierarchical spectrogram transformers," IEEE journal of biomedical and health informatics, vol. 28, no. 3, pp. 1273-1284, 2023, doi: 10.1109/JBHI.2023.3339700.

[25] S. Masri, A. Hasasneh, M. Tami, and C. Tadj, "Exploring the impact of image-based audio representations in classification tasks using vision transformers and explainable AI techniques," Information, vol. 15, no. 12, pp. 751, 2024, doi: 10.3390/info15120751.

[26] S. Li and Y. Sung, "MRBERT: Pre-training of melody and rhythm for automatic music generation," Mathematics, vol. 11, no. 4, pp. 798, 2023, doi: 10.3390/math11040798.

[27] D. V. T. Le, L. Bigo, D. Herremans, and M. Keller, "Natural language processing methods for symbolic music generation and information retrieval: A survey," ACM Computing Surveys, vol. 57, no. 7, pp. 1-40, 2025, doi: 10.1145/3714457.

[28] B. Juarto and A. S. Girsang, "Neural collaborative with sentence BERT for news recommender system," JOIV: International Journal on Informatics Visualization, vol. 5, no. 4, pp. 448-455, 2021, doi: 10.30630/joiv.5.4.678.

[29] C. Hu, X. Sun, H. Dai, H. Zhang, and H. Liu, "Research on log anomaly detection based on sentence-BERT," Electronics, vol. 12, no. 17, pp. 3580, 2023, doi: 10.3390/electronics12173580.

[30] H. Abudukelimu, J. Chen, Y. Liang, A. Abulizi, and A. Yasen, "Symfornet: Application of cross-modal information correspondences based on self-supervision in symbolic music generation," Applied Intelligence, vol. 54, no. 5, pp. 4140-4152, 2024, doi: 10.1007/s10489-024-05335-y.

[31] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, S. Marchand-Maillet, et al., "Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 17, no. 4, pp. 1-23, 2021, doi: 10.1145/3451390.

[32] Y. Zhong, X. Zhang, and Y. Su, "Recommendation method of teaching resources for professional music courses based on knowledge graph," International Journal of High Speed Electronics and Systems, vol. 34, no. 04, Art. no. 2540212, 2025, doi: 10.1142/s0129156425402128.

[33] P. Liu, Y. Cao, and L. Wang, "A multimodal fusion online music education system for universities," Computational Intelligence and Neuroscience, vol. 2022, no. 1, Art. no. 6529110, 2022, doi: 10.1155/2022/6529110.

[34] C. Cayari, "Popular practices for online musicking and performance: Developing creative dispositions for music education and the Internet," Journal of Popular Music Education, vol. 5, no. 3, pp. 295-312, 2021, doi: 10.1386/jpme_00018_1.

[35] J. Hentschel, M. Neuwirth, and M. Rohrmeier, "The annotated mozart sonatas: Score, harmony, and cadence," Transactions of the International Society for Music Information Retrieval, vol. 4, no. 1, pp. 67-80, 2021, doi: 10.5334/TISMIR.63.

[36] J. Hentschel, Y. Rammos, F. C. Moss, M. Neuwirth, and M. Rohrmeier, "An annotated corpus of tonal piano music from the long 19th century," Empirical Musicology Review, vol. 18, no. 1, pp. 84-95, 2024, doi: 10.18061/emr.v18i1.8903.