CLaMP 3: Universal Music Information Retrieval Across Unaligned Modalities and Unseen Languages

Shangda Wu **Zhancheng Guo** **Ruibin Yuan** **Junyan Jiang** **Seungheon Doh**
Gus Xia **Juhan Nam** **Xiaobing Li** **Feng Yu** **Maosong Sun**

Details of authors, correspondence, and affiliations are on Page 9

https://sanderwood.github.io/clamp3

Abstract

CLaMP 3 is a unified framework developed to address challenges of cross-modal and cross-lingual generalization in music information retrieval. Using contrastive learning, it aligns all major music modalities–including sheet music, performance signals, and audio recordings–with multilingual text in a shared representation space, enabling retrieval across unaligned modalities with text as a bridge. It features a multilingual text encoder adaptable to unseen languages, exhibiting strong cross-lingual generalization. Leveraging retrieval-augmented generation, we curated M4-RAG, a web-scale dataset consisting of 2.31 million music-text pairs. This dataset is enriched with detailed metadata that represents a wide array of global musical traditions. To advance future research, we release WikiMT-X, a benchmark comprising 1,000 triplets of sheet music, audio, and richly varied text descriptions. Experiments show that CLaMP 3 achieves state-of-the-art performance on multiple MIR tasks, significantly surpassing previous strong baselines and demonstrating excellent generalization in multimodal and multilingual music contexts.

1 Introduction

Music Information Retrieval (MIR) is a field that aims at developing computational tools for processing, organizing, and accessing music data. A core challenge in MIR is retrieving musical content—whether sheet music, performance signals, or audio recordings—based on natural language queries ("a fast-paced classical piano piece"). This connection enables applications such as automatic music tagging, where models assign genres ("jazz," "folk") or descriptive attributes ("melancholic," "upbeat"), facilitating music organization, search, and recommendation. By integrating NLP methodologies, MIR enables more intuitive access to musical content, making it more interpretable and searchable through text.

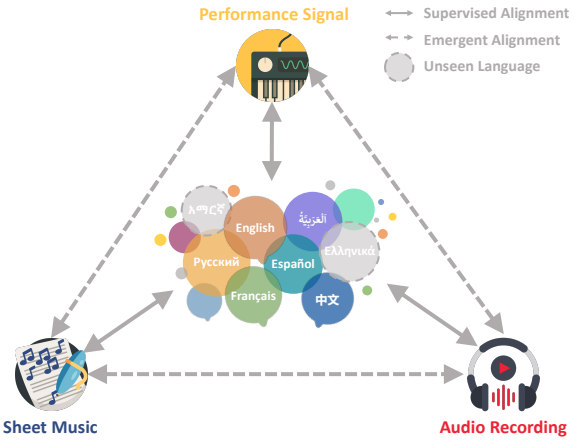


Figure 1: CLaMP 3 demonstrates robust cross-modal and cross-lingual generalization. Supervised alignment (solid arrows) links paired modalities, while emergent alignment (dashed arrows) bridges unaligned ones. A multilingual text encoder enables retrieval in languages unseen (grayed-out bubbles) during alignment.

These capabilities position MIR as a critical bridge between music and language, supporting various applications beyond retrieval and annotation. For instance, cross-modal representations enable text-to-music generation models (Agostinelli et al., 2023; Chen et al., 2024) to create music based on text descriptions. MIR also aids in the automatic evaluation of these models by assessing how closely the generated music aligns with text descriptions or resembles the ground truth (Copet et al., 2023; Retkowski et al., 2024).

Despite these advancements, MIR faces significant challenges in addressing the complexities of ***multimodality*** and ***multilinguality***. Music exists in many forms: sheet music offers human-readable representations for theoretical analysis and education; performance signals (e.g., MIDI) capture timing and dynamics for precise digital editing; and audio recordings serve as the primary medium for listening. While these modalities complement each other, their heterogeneous representational structures complicate unified computational processing.

Findings of the Association for Computational Linguistics: ACL 2025, pages 2605–2625

Adding to this complexity, as a universal medium, music is described in numerous languages, crossing cultural and linguistic boundaries. Musical terminology, descriptions, and cultural references vary significantly between linguistic communities, each bringing its own rich vocabulary and cultural context. To build global and accessible MIR systems, it is essential to process and understand these diverse expressions effectively.

Unfortunately, the development of MIR is limited not only by the lack of music-text pairs but also by the general scarcity of paired data across different musical modalities. As a result, most research focuses on retrieval between specific modality pairs, such as text and audio (Huang et al., 2022; Doh et al., 2024; Zhu et al., 2025) or text and sheet music (Wu et al., 2023a). This narrow focus restricts the potential for cross-modal interactions, preventing a more comprehensive understanding of music. Additionally, existing text data is often short-form, like tags, with few long-form descriptions (Wu et al., 2023b), leading to shallow semantics. These datasets are also predominantly in English (Doh et al., 2023b), with limited representation of other languages, neglecting music's global and multilingual nature.

To tackle these challenges, a unified framework is crucial for aligning musical modalities and bridging linguistic gaps, particularly in the absence of paired training data. Large Language Models (LLMs) present a promising solution by addressing the limitations of text semantics and the scarcity of linguistic diversity in music-text datasets. These models excel at transforming basic metadata into fluent and contextually rich descriptions (Doh et al., 2023a; Bai et al., 2024). Furthermore, their multilingual capabilities allow them to support a wide array of languages (Wu et al., 2024), enhancing semantic depth and enabling more inclusive access across diverse linguistic and cultural contexts.

In this paper, we introduce CLaMP 3, a universal MIR framework that processes music and text while aligning them into a shared representation space. It covers all major music modalities: 1) sheet music, 2) performance signals, and 3) audio recordings, along with 4) multilingual text. Each modality is encoded through its respective feature extractor. To unify these representations, we employ contrastive learning (Sohn, 2016), aligning both musical and textual features. This enables seamless cross-modal retrieval and integration across diverse musical formats and languages.

To address the shortage of paired music-text data, we use Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) to create M4-RAG, a dataset of 2.31 million music-text pairs covering various musical modalities. Starting with basic metadata like song titles and artist names, we retrieve relevant web documents and use an LLM to generate detailed annotations. These annotations include short tags, long descriptions, and multilingual translations, providing rich and diverse information.

In addition, we present WikiMT-X, the first benchmark to align text, audio, and sheet music. It includes 1,000 triplets with diverse text annotations, such as genre labels and detailed long-form descriptions, including background context, musical analysis, general descriptions, and scene depictions. WikiMT-X facilitates evaluation across modalities and semantic perspectives, providing a holistic framework to assess models' ability to align and interpret musical content.

Experiments demonstrate that CLaMP 3 achieves state-of-the-art performance on various MIR tasks, including text-to-audio and text-to-symbolic music retrieval, significantly surpassing all baselines. It also excels in multilingual retrieval, generalizing to languages not present during alignment. By leveraging text as a bridge, CLaMP 3 enables emergent cross-modal retrieval, connecting musical modalities without paired training data.

Overall, this work contributes:

- CLaMP 3 unifies musical modalities and languages in a shared representation space, achieving strong performance on a wide range of MIR tasks and generalizing to unseen languages with emergent cross-modal alignment.
- We curate M4-RAG, a dataset of 2.31 million music-text pairs with diverse annotations, spanning 27 languages and 194 countries, addressing a critical gap in high-quality training data for music and language tasks.
- WikiMT-X links text, audio, and sheet music with 1,000 triplets, offering a first-of-its-kind resource to evaluate models holistically across different modalities and semantic aspects.

To support future research, we have publicly released the complete codebase, pre-trained weights of CLaMP 3, 1.56 million audio-text training pairs, and the WikiMT-X benchmark[1].

[1] https://github.com/sanderwood/clamp3

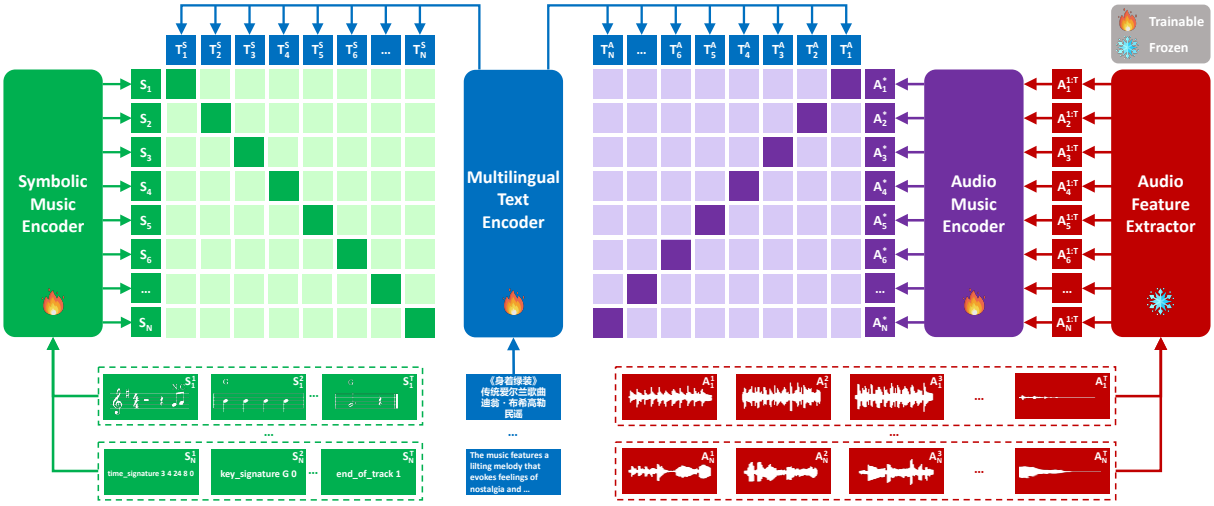


Figure 2: CLaMP 3 uses contrastive learning to align features across modalities. Sheet music and performance signals are segmented into units (bars or MIDI messages) and processed by the symbolic music encoder, while audio is segmented into 5-second clips and processed through the audio feature extractor and audio music encoder. Both symbolic and audio representations are aligned with text representations from the multilingual text encoder.

2 Model

2.1 Training Objective

CLaMP 3's optimization objective is to minimize the InfoNCE loss (Oord et al., 2018), aligning embeddings using contrastive learning:

$$L_{CL} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(z_i^t, z_i^m)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z_i^t, z_j^m)/\tau)}, \quad (1)$$

where z_i^t and z_i^m are text and music embeddings, $\text{sim}(\cdot,\cdot)$ is the similarity function (e.g., dot product), and τ is the temperature parameter. Positive pairs are aligned text-music samples, while negatives are unrelated samples from the same batch.

Inspired by ImageBind (Girdhar et al., 2023), we adopt a multi-stage strategy using text as a bridge to address the lack of paired music data:

Stage 1: The text encoder is first trained to align with one music encoder (e.g., symbolic encoder).

Stage 2: It is then aligned with another music encoder (e.g., audio encoder), freezing the text encoder to prevent representation drift.

Stage 3: The text encoder is unfrozen to refine its alignment with the music encoder from Stage 2.

Stage 4: The text encoder is frozen again to prevent shifts while re-aligning with the Stage 1 music encoder to fix alignment drift from Stage 3.

This strategy minimizes modality interference while mapping all modalities into a shared representation space for effective cross-modal transfer.

2.2 Core Components

CLaMP 3 consists of several transformer-based encoders (Vaswani et al., 2017) for each modality:

Multilingual Text Encoder: The text encoder in CLaMP 3 is based on XLM-R-base (Conneau et al., 2020), a model pre-trained on 2.5 TB of CommonCrawl data across 100 languages. It has 12 layers and a hidden size of 768, enabling strong cross-lingual generalization to unseen languages.

Symbolic Music Encoder: CLaMP 3 uses M3 (Wu et al., 2024), a self-supervised model for encoding symbolic music, including multi-track voice-interleaved ABC notation and lossless MIDI encoding via MIDI Text Format (MTF). M3 segments ABC into bars and MIDI into messages, treating each segment as a patch. The model has 12 encoder layers, a hidden size of 768, and processes up to 512 patches or 32,768 characters per input.

Audio Music Encoder: It is a 12-layer transformer with a 768-dimensional hidden size, trained from scratch for audio processing. This encoder leverages pre-trained features from MERT-v1-95M (Li et al., 2024), where MERT serves as a frozen audio feature extractor. Each 5-second clip is represented by a single embedding, obtained by averaging across all MERT layers and time steps. CLaMP 3 processes up to 128 such embeddings, covering 640 seconds of audio, allowing it to capture high-level audio patterns over extended durations.

All encoders process their outputs through a linear layer, followed by average pooling, to generate a single global semantic feature for each input.

Table 1: Metadata overview for M4-RAG, grouped into basic information, annotations, and translations. In ***Annotations***, *Region* and *Language* are written in English; other fields follow the *Language* specification.

Category	Field	Content	Avg Bytes
Basic	*Title*	Music Title	20.04
	Artists	Artist names	21.97
Annotations	*Region*	Country of origin	20.69
	Language	Document language	7.02
	Genres	Genre list	21.83
	Tags	Keywords/playlists	51.91
	Background	Background context	531.79
	Analysis	Musical analysis	770.29
	Description	General description	591.86
	Scene	Scene depiction	750.92
Translations	*Language*	Translation language	6.38
	Background	Translated background	819.76
	Analysis	Translated analysis	1130.47
	Description	Translated description	888.86
	Scene	Translated scene	1077.07

3 Dataset

In this section, we introduce the M4-RAG dataset for training CLaMP 3 and the WikiMT-X benchmark for evaluation. We start with data sources, followed by the metadata curation process. Then, we summarize dataset statistics like scale and diversity. Finally, we elaborate on the details of the WikiMT-X benchmark.

3.1 Data Sources

The training data for CLaMP 3 is built from both symbolic and audio music datasets, ensuring a rich and diverse foundation for multimodal learning.

The symbolic music data is sourced from WebMusicText (WebMT) (Wu et al., 2023a) with 1.4 million ABC notation files and the Million MIDI Dataset (MMD) (Zeng et al., 2021) with 1.5 million MIDI files. Since symbolic music formats use discrete symbols to represent music, they can be converted into one another, albeit with some information loss. To fully utilize the data, these datasets were unified by converting MMD to ABC and WebMT to MIDI. This process yields 3 million symbolic music files, offering diverse and comprehensive training coverage.

The audio data is collected from online sources, comprising 160 thousand hours of audio from 1.8 million tracks. As CLaMP 3 directly utilizes pre-extracted features, the training data exclusively consists of these precomputed features, leading to substantial savings in both computational resources and time.

3.2 Metadata Curation

Music titles often serve as unique identifiers, enabling the retrieval of rich and detailed descriptions from diverse online sources. When paired with artist names, they further refine searches, pinpointing specific versions or performances and reducing ambiguities caused by covers or adaptations. This distinctive property makes music titles a reliable basis for generating annotations, even in the absence of paired music-text datasets.

To leverage this, we curated M4-RAG (Million-scale Multilingual Music Metadata), a dataset comprising 2.31 million metadata entries. The curation process involved several key steps:

Title Filtering: Entries without titles were excluded, as titles are essential for retrieving meaningful information from the web.

Web Search: Google searches were conducted using titles and, where available, artist names. For each entry, the top 10 search results were collected to ensure diverse and reliable sources.

RAG: Using Qwen2.5-72B (Yang et al., 2024), we generated annotations from the retrieved documents and basic metadata (titles and artist names). The annotations covered the fields in Table 1 under ***Annotations***, with an additional Boolean field indicating if the source material had sufficient information for generating meaningful annotations.

Quality Filtering: Entries were discarded if flagged by the Boolean field for insufficient information, if their format failed to meet the standards outlined in Table 1, or if any fields were left empty.

Postprocessing: To address inconsistencies in the generated annotations, *Region* fields were mapped to recognized countries, while *Description* fields were refined using Qwen to remove identifiable details such as titles and lyrics. Language consistency across long-form fields (*Background*, *Analysis*, *Description*, *Scene*) was verified with fastText (Joulin et al., 2017). Entries with inconsistent languages or languages unsupported by either XLM-R or Qwen were removed, and valid detected languages were recorded in the *Language* field.

Multilingual Translation: To enhance linguistic diversity, a random language supported by both XLM-R and Qwen—different from the original—was selected for each entry, and long-form annotations were translated into it using Qwen.

Prompt and examples of generated annotations are provided in Appendix A.

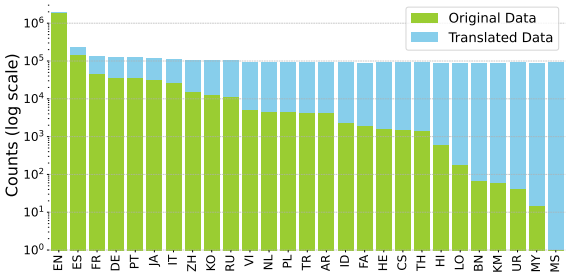


Figure 3: Language distribution of original and translated entries in M4-RAG, covering 27 languages.

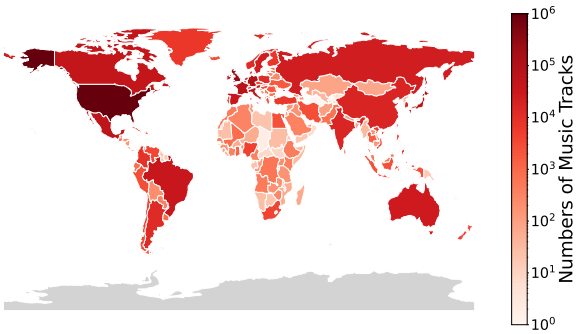


Figure 4: Country-wise distribution of music tracks in M4-RAG, spanning 194 countries.

3.3 Dataset Statistics

Through metadata curation, we obtained M4-RAG, which consists of 2.31 million entries. It includes 0.58 million ABC-text pairs from WebMT, 0.17 million MIDI-text pairs from MMD, and 1.56 million audio-text pairs.

Each metadata entry includes both short-form annotations, such as genres and tags, and detailed long-form descriptions. As summarized in Table 1, the long-form descriptions account for the majority of the dataset, providing extensive semantic details from multiple perspectives.

M4-RAG spans 27 languages, with the original metadata predominantly in English, as shown in Fig. 3. To address this imbalance, translations were added to the long-form descriptions, greatly boosting non-English data. This was particularly impactful for low-resource languages, such as Malay and Burmese, where most data depends on translations, greatly enhancing their representation.

In terms of geographic coverage, M4-RAG incorporates music from 194 countries. Fig. 4 illustrates contributions from both major music-producing nations and less-represented regions. This global reach ensures the dataset reflects a diverse range of musical traditions and styles from across the world.

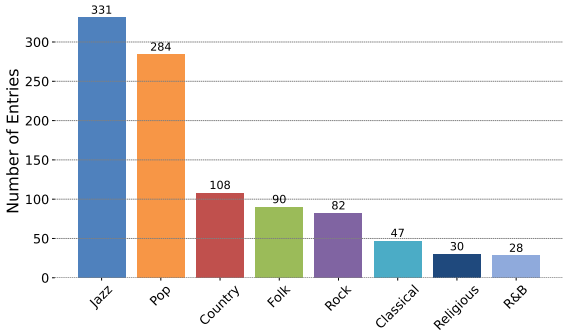


Figure 5: Genre distribution of the WikiMT-X dataset.

3.4 Benchmark Dataset

WikiMT-X (WikiMusicText-eXtended) extends WikiMT (Wu et al., 2023a), focusing on 20th-century Western music with 1,000 entries, each with sheet music, audio, and detailed metadata.

The original WikiMT dataset had the following drawbacks: 1) the text was sourced from Wikipedia, mainly focused on background information with limited semantic diversity; 2) the absence of audio data severely restricted the evaluation scope; and 3) the genre labels were obtained through keyword matching, resulting in relatively low accuracy and reducing the reliability of the dataset.

To address these deficiencies, WikiMT-X made the following improvements:

- We used llama-3.1-sonar-large-128k-online[2] (Dubey et al., 2024), feeding it sheet music with titles, artist names, and lyrics. It retrieved relevant web pages and summarized them into background, analysis, description, and scene.
- We manually matched sheet music with audio recordings retrieved from YouTube and removed 10 identified duplicates.
- We reorganized genre categories based on data distribution and re-annotated labels.

These enhancements make WikiMT-X useful for multimodal MIR research tasks, assessing models' capabilities in handling text annotations of diverse semantic types, and classifying music across modalities using genre labels.

Appendix B presents detailed objective and human evaluations of WikiMT-X annotation quality. In addition, Appendix C provides *t*-SNE visualizations of CLaMP 3 embeddings on WikiMT-X, showing modality, language, and semantic distributions in the shared representation space.

[2] https://www.perplexity.ai

Table 2: Results for English text-to-music retrieval on several benchmarks: WikiMT and MidiCaps have 1,010 pairs, Song Describer Dataset (SDD) has 706 audio and 1,106 captions, and MusicCaps-Remake (MC-R) contains 2,777 pairs. MC-R prevents data leakage by using full-length audio and rewritten captions from AudioSet's evaluation set.

Model	Symbolic Benchmarks		WikiMT-X (Sheet Music)			
	WikiMT	*MidiCaps*	*Background*	*Analysis*	*Description*	*Scene*
CLaMP	0.2561	0.1236	0.2122	0.1345	0.0306	0.0426
CLaMP 2	0.3438	0.2695	0.3024	0.2374	0.0418	0.0838
CLaMP 3^{c2}_{sa}	**0.4498**	**0.2826**	**0.4028**	**0.3382**	0.0835	**0.1512**
CLaMP 3_{saas}	0.3555	0.1798	0.3301	0.2758	**0.1274**	0.1500
Model	**Audio Benchmarks**		**WikiMT-X (Audio)**			
	SDD	*MC-R*	*Background*	*Analysis*	*Description*	*Scene*
CLAP	0.1310	0.0657	0.0598	0.0429	0.0318	0.0218
TTMR++	0.1437	**0.1248**	0.1119	0.0833	0.0584	0.0301
CLaMP 3^{c2}_{sa}	0.1612	0.0959	0.1180	0.1206	0.0639	0.0619
CLaMP 3_{saas}	**0.1985**	0.1177	**0.2017**	**0.1711**	**0.0988**	**0.0963**

4 Experiments

This section evaluates CLaMP 3 on retrieval tasks, comparing it to state-of-the-art baselines. We present results for the two best-performing CLaMP 3 variants—one for symbolic music and one for audio. A full retrieval comparison of all variants can be found in Appendix D, and classification results are available in Appendix E.

4.1 Settings

Both symbolic music and audio alignments were trained for up to 100 epochs on 8 NVIDIA H800 GPUs. Symbolic music alignment required 4 days with a learning rate of 5e-5 and a batch size of 1024. Audio alignment took 1 day with a learning rate of 1e-5 and a batch size of 2048.

M4-RAG was divided into 99% for training and 1% for validation. During training, metadata information was randomly selected to form text inputs. Mixed-precision (Micikevicius et al., 2018), AdamW optimizer (Loshchilov and Hutter, 2019), and a 1,000-step warm-up (Goyal et al., 2017) were used to enhance efficiency.

Following the training strategy in Sec. 2.1, we explored various modality alignment orders for symbolic and audio modalities, and present the two top-performing variants below:

CLaMP 3_{saas}: Optimized for audio, this model follows the full multi-stage alignment: symbolic → audio → audio → symbolic.

CLaMP 3^{c2}_{sa}: Optimized for symbolic, this model starts from CLaMP 2-initialized text and symbolic encoders, followed by two stages: the text encoder is jointly trained with the symbolic encoder, then frozen to align with the audio encoder.

4.2 English Text-to-Music Retrieval

We evaluated retrieval performance using Mean Reciprocal Rank (MRR), which measures the inverse of the rank of the paired item, across all tasks.

For symbolic music retrieval, we compared CLaMP 3 with CLaMP 2 (Wu et al., 2024) and CLaMP (Wu et al., 2023a) on WikiMT (using ABC notation) and MidiCaps (Melechovsky et al., 2024) (using MIDI). For audio retrieval, we evaluated CLaMP 3 against state-of-the-art models CLAP (Wu et al., 2023b) and TTMR++ (Doh et al., 2024) on the Song Describer Dataset (SDD) (Manco et al., 2023) and MusicCaps-Remake (MC-R) (Agostinelli et al., 2023), which addresses data leakage by using full-length audio and rewritten captions (see Appendix F) from AudioSet's evaluation set (Gemmeke et al., 2017). In addition, we tested all models on WikiMT-X to evaluate their performance across varying semantic perspectives.

As shown in Table 2, CLaMP 3 achieved significant improvements over its predecessors and baseline models across both symbolic and audio retrieval tasks. For symbolic music retrieval, CLaMP 3^{c2}_{sa} achieved MRR scores of 0.4498 on WikiMT and 0.2826 on MidiCaps, clearly outperforming both CLaMP 2 and CLaMP, despite using only half the training data. This improvement can be attributed to the high-quality, richly annotated M4-RAG dataset. Similarly, CLaMP 3_{saas}, though optimized for audio retrieval, exceeded CLaMP by a notable margin on symbolic benchmarks and performed comparably to CLaMP 2 on WikiMT. These results demonstrate that our multi-stage training approach effectively preserves performance on modalities that were not explicitly optimized.

Table 3: Results for multilingual text-to-music retrieval on translated WikiMT-X background annotations. Languages marked with asterisks were not included in the M4-RAG training data. The BLEU scores below each language are calculated by back-translating the text with the SeamlessM4T model and comparing it to the original English text.

Model	**ru** *49.69*	**fr** *55.50*	**es** *62.82*	**ar** *53.38*	**zh** *39.58*	**fi*** *39.19*	**el*** *55.55*	**ta*** *40.07*	**kk*** *36.57*	**am*** *56.08*
ABC Notation										
CLaMP 2	0.2668	0.2968	0.2934	0.2298	0.1646	0.2795	0.2410	0.0915	0.2543	0.1237
CLaMP 3^{c2}_{sa}	**0.3614**	**0.3949**	**0.3921**	**0.3155**	**0.2373**	**0.3524**	**0.3226**	**0.1415**	**0.3397**	**0.1871**
CLaMP 3_{saas}	0.2918	0.3214	0.3239	0.2789	0.2358	0.2919	0.2681	0.1246	0.2703	0.1139
MIDI										
CLaMP 2	0.1271	0.1414	0.1452	0.1113	0.0749	0.1438	0.1087	0.0466	0.1079	0.0616
CLaMP 3^{c2}_{sa}	**0.1921**	**0.2101**	**0.2137**	**0.1681**	**0.1316**	**0.2019**	**0.1702**	**0.0804**	**0.1765**	**0.1039**
CLaMP 3_{saas}	0.1165	0.1319	0.1330	0.1141	0.0937	0.1245	0.1143	0.0601	0.1104	0.0544
Audio										
CLaMP 3^{c2}_{sa}	0.1068	0.1150	0.1202	0.0981	0.0877	0.1112	0.1014	0.0720	0.1005	**0.0681**
CLaMP 3_{saas}	**0.1788**	**0.1980**	**0.1962**	**0.1665**	**0.1459**	**0.1770**	**0.1736**	**0.0945**	**0.1561**	0.0675

Beyond symbolic music retrieval, CLaMP 3 also achieved notable performances in audio retrieval. Both variants—CLaMP 3^{c2}_{sa} and CLaMP 3_{saas}—consistently outperformed CLAP, with CLaMP 3_{saas} standing out. It achieved the highest MRR of 0.1985 on SDD, marking a substantial improvement over TTMR++ (0.1437) and CLAP (0.1310). While TTMR++ performed well on MC-R (0.1248), its results on the original MusicCaps dataset are abnormally higher (see Table 13), likely because it was trained on half of MusicCaps' original music-text pairs. This training overlap suggests that indirect data leakage affects its performance, even when evaluated on MC-R.

CLaMP 3's strong performance extends to WikiMT-X, with both variants outperforming baselines across all four semantic categories. In *Background* and *Analysis*, where texts provide rich cultural or technical details, CLaMP 3^{c2}_{sa} and CLaMP 3_{saas} excelled, achieving MRRs of 0.4028 and 0.3382 (sheet music) and 0.2017 and 0.1711 (audio). *Description* and *Scene*, however, are much harder to retrieve because they are less specific and semantically sparse. *Description* excludes explicit identifiers like titles or artist names, while *Scene* focuses on abstract, visualized scenario depictions (rather than the music itself), both of which make retrieval more difficult. Even so, CLaMP 3 performed notably better, with CLaMP 3_{saas} scoring 0.0988 (*Description*) and 0.0963 (*Scene*) in audio, compared to TTMR++ (0.0584, 0.0301). This improvement stems from M4-RAG's diverse annotations, which better equip CLaMP 3 to retrieve abstract, semantically sparse texts compared to baseline models trained on less diverse data.

4.3 Multilingual Text-to-Music Retrieval

Currently, no non-English music-text benchmarks exist, making multilingual evaluation challenging. To address this, we used SeamlessM4T (Barrault et al., 2023) to translate WikiMT-X background annotations into multiple languages. To account for translation noise, BLEU scores (Papineni et al., 2002) were calculated by comparing original texts with back-translations. The translated annotations were then used for retrieval of matching ABC notation, MIDI (from ABC), and audio files.

We carefully selected ten languages to ensure diversity in linguistic families, scripts, regions, and resource levels. Five UN official languages were chosen from those included in M4-RAG as they represent different cultures and regions with global significance. The other five, marked with asterisks in Table 3, come from different linguistic families with distinct scripts and minimal vocabulary overlap, specifically to test CLaMP 3's generalization to languages unseen in music-text alignment.

To the best of our knowledge, apart from CLaMP 3, CLaMP 2 is the only multilingual MIR model, but it is limited to symbolic music. No baselines exist for multilingual audio retrieval, as models like CLAP and TTMR++ are restricted to English.

CLaMP 3's two variants differ in their language exposure. CLaMP 3^{c2}_{sa} initializes its text and symbolic music encoders from CLaMP 2, which was pre-trained on symbolic-text alignment across all XLM-R-supported languages, giving it prior exposure to all languages in Table 3. In contrast, CLaMP 3_{saas} has never aligned music data with the languages marked with asterisks, demonstrating true cross-lingual generalization in its performance.

Table 4: Results for emergent cross-modal retrieval on WikiMT-X pairings across different musical modalities. **S**: Sheet Music (ABC notation), **P**: Performance Signals (MIDI, converted from ABC), **A**: Audio recordings.

Model	S→P	S→A	P→S	P→A	A→S	A→P
CLaMP 2	**0.5138**	-	0.4480	-	-	-
CLaMP 3^{c2}_{sa}	0.4547	0.0543	**0.5293**	0.0313	**0.0492**	**0.0383**
CLaMP 3_{saas}	0.3262	**0.0578**	0.3146	**0.0397**	0.0410	0.0303

Table 3 shows that CLaMP 3 demonstrates strong cross-lingual generalization in both symbolic music and audio retrieval tasks. For symbolic music retrieval, CLaMP 3^{c2}_{sa} clearly outperforms CLaMP 2 on all languages, including those not in M4-RAG, showing that full language coverage during training is not necessary for improved multilingual retrieval. Meanwhile, CLaMP 3_{saas}, without any prior alignment between these languages and music or specific optimization for symbolic music tasks, matches CLaMP 2's performance on MIDI and surpasses it on ABC notation. This indicates that CLaMP 3_{saas} achieves true cross-lingual generalization on unseen languages.

In audio retrieval, CLaMP 3_{saas} performed well on languages it had never seen during alignment. For instance, it outperformed CLaMP 3^{c2}_{sa} on Finnish (0.1770 vs. 0.1112), Greek (0.1736 vs. 0.1014), and Kazakh (0.1561 vs. 0.1005), even though CLaMP 3^{c2}_{sa} had indirect exposure to these languages during CLaMP 2 pre-training. Notably, even for its weakest unseen language, Amharic (0.0675), CLaMP 3_{saas} outperformed CLAP's performance on English text (0.0598). This suggests that prior exposure to a language is not necessary for achieving strong audio retrieval performance.

The ability to retrieve languages beyond the training data stems from XLM-R's cross-lingual semantics and the universal representations of CLaMP 3's music encoders. This enables the model to handle low-resource languages and even generalize to unseen ones, enhancing its inclusivity and versatility for global MIR.

4.4 Emergent Cross-Modal Retrieval

Emergent cross-modal retrieval assesses a model's ability to align and retrieve musical content across modalities without explicit alignment training, showcasing its capacity to generalize to unaligned modalities. Table 4 reports results for all possible retrieval directions between ABC notation, MIDI, and audio data.

CLaMP 3 significantly advances cross-modal retrieval by supporting both symbolic and audio modalities, addressing a key limitation of CLaMP 2. While CLaMP 2 excels in symbolic tasks (S→P: 0.5138, P→S: 0.4480) without explicit alignment between ABC and MIDI, it cannot retrieve between symbolic and audio modalities.

In contrast, CLaMP 3^{c2}_{sa} not only achieves state-of-the-art performance on symbolic music tasks like P→S (0.5293) but also enables emergent retrieval between symbolic music and audio. Similarly, CLaMP 3_{saas}, optimized for audio retrieval, achieves meaningful results on new tasks such as S→A (0.0578) and P→A (0.0397), demonstrating its ability to unify symbolic and audio modalities in a shared representation space.

While audio retrieval is inherently more challenging due to the continuous nature of audio signals, all directions achieve MRR scores well above the random baseline of 0.0075. Nonetheless, further optimization is required to reduce the performance gap between symbolic and audio retrieval.

5 Conclusions

In this paper, we introduced CLaMP 3, a unified MIR framework that aligns sheet music, performance signals, audio, and multilingual text using contrastive learning. CLaMP 3 demonstrates strong cross-modal and cross-lingual generalization, effectively handling unaligned modalities and unseen languages during training.

To address the lack of high-quality datasets, we curated M4-RAG, a collection of 2.31 million music-text pairs spanning 27 languages and 194 countries. We also released WikiMT-X, the first benchmark combining text, sheet music, and audio for comprehensive evaluation.

Our experiments show that CLaMP 3 achieves state-of-the-art performance in both symbolic and audio retrieval, excels in multilingual tasks, and enables retrieval across unaligned musical modalities. These results demonstrate its flexibility and the effectiveness of its shared representation space.

To conclude, CLaMP 3 sets a new standard in multimodal and multilingual MIR, demonstrating robust cross-modal and cross-lingual generalization. By releasing the CLaMP 3 model, M4-RAG dataset, and WikiMT-X benchmark, we provide resources to support future research in MIR and music generation across languages and modalities.

6 Limitations

Although CLaMP 3 attains state-of-the-art performance across modalities and languages, showing cross-modal and cross-lingual generalization, this work has several limitations that need to be addressed for further advancements in MIR.

First, while contrastive learning has advanced multimodal information retrieval, it struggles to capture the temporal dynamics of music. This is because such models typically use a single global representation to store the entire semantic content of a piece of music, making them insensitive to temporal dynamics. For example, in Beethoven's Symphony No. 5, the iconic four-note motif develops throughout the piece, yet current systems often miss this context. Addressing this requires moving beyond contrastive learning to incorporate temporal modeling, enabling systems to better capture nuances and deliver more context-aware and accurate retrieval.

Second, although Table 3 indicates that while CLaMP 3 can generalize to languages beyond music-text alignment, the multilingual text-to-music retrieval evaluation in it heavily relies on translation models due to the lack of native multilingual benchmarks. The translation quality varies significantly across languages, which introduces noise and reduces the reliability of evaluations. Developing native multilingual benchmarks is the primary and almost indispensable solution to achieve more accurate and fair assessments of model performance.

Finally, as shown in Table 4, the alignment between audio and symbolic modalities, though showing emergent capabilities with performance far above random, remains relatively weak. Addressing this limitation requires collecting paired data for supervised alignment and leveraging text as a bridging modality to further enhance connections between different musical modalities.

Authors

Shangda Wu[1], *shangda@mail.ccom.edu.cn*
Zhancheng Guo[1], *23a053@mail.ccom.edu.cn*
Ruibin Yuan[2], *ryuanab@connect.ust.hk*
Junyan Jiang[3, 4], *jj2731@nyu.edu*
Seungheon Doh[5], *seungheondoh@kaist.ac.kr*
Gus Xia[3, 4], *Gus.Xia@mbzuai.ac.ae*
Juhan Nam[5], *juhan.nam@kaist.ac.kr*
Xiaobing Li[1], *lxiaobing@ccom.edu.cn*
Feng Yu[1], *yufengai@ccom.edu.cn*

Correspondence

Maosong Sun[1, 6], *sms@tsinghua.edu.cn*

Affiliations

[1]Central Conservatory of Music
[2]Hong Kong University of Science and Technology
[3]New York University Shanghai
[4]Mohamed bin Zayed University of Artificial Intelligence
[5]Korea Advanced Institute of Science and Technology
[6]Tsinghua University

Acknowlegdements

This work was supported by the following funding sources: Special Program of National Natural Science Foundation of China (Grant No. T2341003), Advanced Discipline Construction Project of Beijing Universities, Major Program of National Social Science Fund of China (Grant No. 21ZD19), and the National Culture and Tourism Technological Innovation Engineering Project (Research and Application of 3D Music).

In addition, we thank Flaticon[3] for icons used in Fig. 1 and Fig. 2, Yusong Wu (University of Montreal) for helping us understand CLAP in detail, and Monan Zhou (Central Conservatory of Music) for assisting with WikiMT-X data processing.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D Plumbley, Woon-Seng Gan, and Jianfeng Chen. 2024. Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models. *arXiv preprint arXiv:2411.18953*.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The mtg-jamendo dataset for automatic music tagging. ICML.

[3]https://www.flaticon.com

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 1206–1210. IEEE.

Yi-Hui Chou, I Chen, Chin-Jui Chang, Joann Ching, Yi-Hsuan Yang, et al. 2021. Midibert-piano: large-scale pre-training for symbolic music understanding. *arXiv preprint arXiv:2107.05223*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Seungheon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023a. Lp-musiccaps: Llm-based pseudo music captioning. In *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, pages 409–416.

Seungheon Doh, Minhee Lee, Dasaem Jeong, and Juhan Nam. 2024. Enriching music descriptions with A finetuned-llm and metadata for text-to-music retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 826–830. IEEE.

Seungheon Doh, Minz Won, Keunwoo Choi, and Juhan Nam. 2023b. Toward universal text-to-music retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lucas Ferreira and Jim Whitehead. 2019. Learning to generate music with sentiment. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, pages 384–390.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 776–780. IEEE.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15180–15190. IEEE.

Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. 2022. Mulan: A joint embedding of music audio and natural language. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, pages 559–566.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomás Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 427–431. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger B. Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. 2024. MERT: acoustic music understanding model with large-scale self-supervised training. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, et al. 2023. The song describer dataset: a corpus of audio captions for music-and-language evaluation. *arXiv preprint arXiv:2311.10057*.

Jan Melechovsky, Abhinaba Roy, and Dorien Herremans. 2024. Midicaps–a large-scale midi dataset with text captions. *arXiv preprint arXiv:2406.02255*.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Jan Retkowski, Jakub Stępniak, and Mateusz Modrzejewski. 2024. Frechet music distance: A metric for generative symbolic music evaluation. *arXiv preprint arXiv:2412.07948*.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1849–1857.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Shangda Wu, Yashan Wang, Ruibin Yuan, Zhancheng Guo, Xu Tan, Ge Zhang, Monan Zhou, Jing Chen, Xuefeng Mu, Yuejie Gao, et al. 2024. Clamp 2: Multimodal music information retrieval across 101 languages using large language models. *arXiv preprint arXiv:2410.13267*.

Shangda Wu, Dingyao Yu, Xu Tan, and Maosong Sun. 2023a. Clamp: Contrastive language-music pre-training for cross-modal symbolic music information retrieval. In *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, pages 157–165.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023b. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Le Zhuo, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, Ningzhi Wang, Chenghua Lin, Emmanouil Benetos, Anton Ragni, Norbert Gyenge, Roger B. Dannenberg, Wenhu Chen, Gus Xia, Wei Xue, Si Liu, Shi Wang, Ruibo Liu, Yike Guo, and Jie Fu. 2023. MARBLE: music audio representation benchmark for universal evaluation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. 2021. Musicbert: Symbolic music understanding with large-scale pre-training. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 791–800. Association for Computational Linguistics.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Wei Tan, and Xie Chen. 2025. Muq: Self-supervised music representation learning with mel residual vector quantization. *arXiv preprint arXiv:2501.01108*.

A Prompt and Examples

Your task is to provide a detailed set of metadata for a music piece in JSON format based on the input provided. The input includes the following fields:

```json
{
  "title": "Music title [in any language] (string)",
  "artists": ["List of singer/performer/composer/lyricist names [in any language] (array of strings) (optional)"],
  "search_results": [
    {
      "title": "Search result title [in any language] (string)",
      "text": "Search result content [in any language] (string)"
    },
    {
      "title": "Search result title [in any language] (string)",
      "text": "Search result content [in any language] (string)"
    }
  ]
}
```

You MUST USE THE SEARCH RESULTS THOROUGHLY to GATHER AS MUCH RELEVANT INFORMATION AS POSSIBLE. Ensure every field in your output is comprehensive, accurate, and formatted according to JSON standards. Adhere strictly to the following JSON format in your response, without adding explanations or comments:

```json
{
  "sufficient_information": "Boolean value indicating whether the search results provide enough information to summarize metadata for the music (true/false)",
  "title": "Music title [in the original language] (string)",
  "artists": ["List of singer/performer/composer/lyricist names [in the original language] (array of strings)"],
  "region": "The country or geographical region associated with the music [in English] (string)",
  "language": "Language in which the search results are primarily written [in English] (string)",
  "genres": ["List of genres [in the specified language] (array of strings)"],
  "tags": ["List of tags/keywords/playlists/moods [in the specified language] (array of strings)"],
  "background": "Background information and fun facts about the music [in the specified language] (string)",
  "analysis": "In-depth, thorough, and academic-level musical analysis [in the specified language] (string)",
  "description": "A detailed de-identified description without identifying specifics [in the specified language] (string)",
  "scene": "A vivid textual description of the visual scene where this music is used as a soundtrack or ideally played [in the specified language] (string)"
}
```

Important Notes:
1. If you find that the title or artist information is inaccurate, fill in the correct information based on the search results.
2. Leave fields empty ("" or []) if specific information (e.g., region) is not provided in the search results.
3. Language Requirements:
   - The 'title' and 'artists' fields must be written in their original language, as applicable.
   - The 'region' and 'language' fields must be written in English.
   - All other fields must be written in the language specified in the 'language' field.
4. Background Field:
   - The 'background' field should focus on the creation, cultural, and social context of the piece. This may include the origin of the work, the cultural or social impact, the artist's role in shaping the piece, and any notable reception or influence the work has had.
5. Analysis Field:
   - The 'analysis' field should provide a detailed breakdown of the musical elements. This may include the structure and form, melody and harmony, rhythm and meter, instrumentation, emotional tone, style and genre, and any influences the piece draws from or has had on later works.
6. Description Field:
   - Create a detailed yet de-identified summary of the music's key features and style. AVOID MENTIONING IDENTIFIABLE DETAILS (e.g., the title, artist names, or lyrics).
7. Scene Field:
   - If it's a soundtrack, you MUST DESCRIBE THE EXACT MOST REPRESENTATIVE SCENE IT ACCOMPANIES, providing rich visual and sensory details. Include the setting (such as location, time of day, and lighting), characters (their appearance, emotions, and actions), objects in the scene, and any significant events or emotional tones that align with the music.
   - If it's not a soundtrack, describe detailed visual elements that match the music, creating an evocative image where the music naturally enhances the visual and emotional atmosphere.

Figure 6: The metadata generation prompt was used for constructing the M4-RAG dataset. This prompt outlines the required JSON structure for describing music metadata comprehensively, including fields for title, artists, region, language, genres, tags, background context, musical analysis, general description, and visual scene. Detailed instructions and formatting requirements are provided to ensure high-quality and consistent metadata extraction from search results. Based on our experience, we recommend adding the requirement to the prompt that *Region* and *Language* be output in accordance with ISO standards, which can reduce the need for post-processing.

```
{
  "id": "---aL9TdeI4",
  "title": "Mairi's Wedding",
  "artists": [
    "Noel McLoughlin"
  ],
  "region": "United Kingdom of Great Britain and Northern Ireland",
  "language": "English",
  "genres": [
    "Folk",
    "Traditional"
  ],
  "tags": [
    "Scottish",
    "Wedding",
    "Traditional",
    "Folk",
    "Celtic"
  ],
  "background": "Mairi's Wedding (also known as Marie's Wedding, the Lewis Bridal Song, or Mairi Bhan) is a Scottish folk song originally written in Gaelic by Johnny Bannerman for Mary McNiven. Written using a traditional Scottish tune, it was first played for McNiven in 1935 at the Old Highlanders Institute in Glasgow's Elmbank Street. Hugh S. Roberton translated the Gaelic version into English in 1936. The song has since become a popular traditional Scottish folk song, often performed by various artists including The High Kings, The Clancy Brothers & Tommy Makem, and Noel McLoughlin.",
  "analysis": "Mairi's Wedding is a lively and upbeat Scottish folk song with a traditional reel structure. The song is in a major key, which contributes to its celebratory and joyful mood. The lyrics describe the journey of guests to Mairi's wedding, the bride's beauty, and a toast to her future happiness. The chorus is particularly rhythmic and danceable, with the repeated phrase 'Step we gaily on we go' encouraging participation and movement. The instrumentation typically includes acoustic instruments such as the guitar, banjo, and fiddle, which add to the song's traditional and authentic feel. The song's structure, with its alternating verses and choruses, is typical of many folk songs and helps to maintain engagement and energy throughout the performance.",
  "description": "This is a lively and upbeat traditional folk piece that captures the spirit of a joyful celebration. It features a repetitive and danceable chorus, accompanied by cheerful and rhythmic instrumentation. The lyrics paint a vivid picture of a wedding day, with guests traveling over hills and through towns to join in the festivities. The music emphasizes the beauty and charm of the occasion, as well as the hope for a prosperous and happy future. The melody is catchy and the overall mood is one of happiness and community. The piece is rooted in Scottish and Celtic traditions, reflecting the rich cultural heritage of the region.",
  "scene": "The scene is set in a picturesque Scottish village on a sunny morning. The air is crisp and the sky is a clear blue. Villagers are seen walking through the winding streets, their steps quick and lively. They carry flowers and baskets, their faces beaming with excitement. The path they follow leads to a small, rustic church adorned with fresh greenery and wildflowers. As they approach, the sound of a fiddle and a bodhrán can be heard, signaling the start of the wedding festivities. Inside the church, the bride, Mairi, stands nervously but beautifully, her cheeks flushed with happiness. The guests, arm in arm, step gaily to the music, their feet moving in time to the rhythmic beat. The atmosphere is one of joy and celebration, with everyone eager to partake in the wedding of Mairi and her beloved.",
  "translations": {
    "language": "Vietnamese",
    "background": "Bài hát \"Đám Cưới Mairi\" (còn được biết đến với tên gọi \"Đám Cưới Marie\", \"Bài Ca Cô Dâu Lewis\" hoặc \"Mairi Bhan\") là một bài hát dân gian Scotland do Johnny Bannerman sáng tác bằng tiếng Gaelic dành cho Mary McNiven. Sử dụng giai điệu truyền thống Scotland, bài hát được trình diễn lần đầu cho McNiven vào năm 1935 tại Old Highlanders Institute trên đường Elmbank, Glasgow. Năm 1936, Hugh S. Roberton đã dịch phiên bản tiếng Gaelic sang tiếng Anh. Từ đó, bài hát đã trở thành một bài hát dân gian Scotland truyền thống phổ biến, thường được biểu diễn bởi nhiều nghệ sĩ bao gồm The High Kings, The Clancy Brothers & Tommy Makem, và Noel McLoughlin.",
    "analysis": "\"Bài hát Cưới Hỏi của Mairi\" là một bài hát dân gian Scotland sôi động và vui tươi, có cấu trúc theo kiểu điệu reel truyền thống. Bài hát được sáng tác theo cung trưởng, điều này góp phần tạo nên không khí lễ hội và hạnh phúc. Lời bài hát mô tả chuyến đi của khách dự tiệc đến đám cưới của Mairi, vẻ đẹp của cô dâu, và lời chúc mừng hạnh phúc tương lai của cô. Điệp khúc đặc biệt nhịp nhàng và dễ nhảy múa, với câu lặp lại \"Bước chân vui tươi, chúng ta tiến bước\" khích lệ sự tham gia và vận động. Phần đệm đàn thường bao gồm các nhạc cụ acoustic như guitar, banjo, và fiddle, làm tăng thêm vẻ truyền thống và chân thực của bài hát. Cấu trúc bài hát, với các đoạn thơ và điệp khúc xen kẽ, là đặc trưng của nhiều bài hát dân gian, giúp duy trì sự hấp dẫn và sức sống suốt buổi trình diễn.",
    "description": "Đây là một tác phẩm dân gian truyền thống sôi động và vui tươi, nắm bắt tinh thần của một buổi lễ ăn mừng đầy hân hoan. Bài hát có phần điệp khúc lặp lại và dễ nhảy múa, được acompanh bởi những âm thanh nhạc cụ vui vẻ và có tiết tấu. Lời bài hát vẽ nên bức tranh sinh động về ngày cưới, với khách mời vượt qua những ngọn đồi và đi xuyên qua các thị trấn để tham gia vào buổi tiệc. Nhạc phẩm nhấn mạnh vẻ đẹp và sự quyến rũ của dịp lễ, cũng như niềm hy vọng về một tương lai thịnh vượng và hạnh phúc. Điệu nhạc dễ nhớ và tâm trạng chung là niềm vui và sự đoàn kết. Bài hát có nguồn gốc từ truyền thống Scotland và Celtic, phản ánh di sản văn hóa phong phú của khu vực.",
    "scene": "Bối cảnh diễn ra trong một ngôi làng xinh đẹp của Scotland vào một buổi sáng nắng đẹp. Không khí trong lành và bầu trời trong xanh. Người dân trong làng được thấy đang đi qua những con đường uốn lượn, bước chân nhanh nhẹn và vui tươi. Họ mang theo hoa và giỏ đồ, khuôn mặt rạng rỡ với niềm hân hoan. Đường họ đi dẫn đến một ngôi nhà thờ nhỏ, cổ kính, được trang trí bằng cây cỏ tươi mới và hoa dại. Khi họ tiến gần, âm thanh của cây đàn fiddle và trống bodhrán vang lên, báo hiệu lễ cưới sắp bắt đầu. Bên trong nhà thờ, cô dâu Mairi đứng ở đó, vừa lo lắng vừa xinh đẹp, má ửng hồng hạnh phúc. Khách mời, tay trong tay, bước nhảy nhót theo nhạc, chân di chuyển nhịp nhàng theo tiếng trống. Không khí tràn đầy niềm vui và sự chúc tụng, mọi người đều mong muốn tham gia vào lễ cưới của Mairi và người yêu dấu của cô."
  }
}
```

```
{
  "id": "Vj4gbky1QCU",
  "title": "'Deed I Do",
  "artists": [
    "Fred Rose",
    "Walter Hirsch"
  ],
  "genre": "Jazz",
  "background": "Originally composed in the early 20th century, 'Deed I Do' has been performed by numerous artists and remains a staple in jazz and pop standards. The song's lyrics express themes of love and devotion, often interpreted through various musical styles.",
  "analysis": "'Deed I Do' features a simple yet effective harmonic structure that allows for expressive vocal interpretations. The song typically follows a AABA form, common in jazz standards, where the A sections present the main theme and the B section offers a contrasting idea. The melody is characterized by its lyrical phrasing, which provides opportunities for improvisation and personal expression by performers. The use of syncopation and swing rhythm enhances its playful yet romantic feel, making it a favorite among jazz musicians. The lyrics convey a deep emotional connection, often evoking feelings of nostalgia and longing.",
  "description": "The music is characterized by a smooth, flowing melody that captures the essence of romantic yearning. It features lush harmonies and a moderate tempo that invites listeners to engage with its sentimental lyrics. Instrumentation typically includes piano, bass, and light percussion, creating an intimate atmosphere that complements the vocal line. The overall sound is warm and inviting, making it suitable for both casual listening and more formal performances.",
  "scene": "The music plays softly in a dimly lit jazz club, where patrons are seated at small tables adorned with flickering candles. The air is filled with the rich aroma of coffee and dessert as couples lean in closer to hear the tender notes. A singer stands center stage under a spotlight, their voice smooth and emotive, capturing the attention of an audience lost in the moment. Outside, rain gently taps against the window, adding to the cozy ambiance as the singer expresses heartfelt sentiments about love and devotion.",
  "leadsheet": "X:1\nT:\"Deed I Do\nT:'Deed I Do\nC:Fred Rose\nL:1/8\nM:2/2\nI:linebreak $\nK:C\nV:1 treble nm=\"Voice\"\n%%MIDI program 52\nV:1\n\"Cmaj7\" A c3- c4 |\"Gm7\" A G3-\"C7\" G4 |\"Fmaj7\" E D3- D4 |\"Fm\" E D3- D4 |\"Em7\" E G3-\"A7\" G4 | %5\nw: Do I *|want you? *|Oh my, *|do I? *|Hon- ey *|\n\"D7\" C4\"G7\" D4 |\"Cmaj7\" C8- |\"Dm7\" C4\"G7\" z4 |\"Cmaj7\" A c3- c4 |\"Gm7\" A G3-\"C7\" G4 | %10\nw: 'deed I|do.||Do I *|need you? *|\n\"Fmaj7\" E D3- D4 |\"Fm\" E D3- D4 |\"Em7\" E G3-\"A7\" G4 |\"D7\" C4\"G7\" D4 |\"Cmaj7\" C8- |\"Gm7\" C4\"C7\" z4 | %16\nw: Oh my, *|do I? *|Hon- ey *|'deed I|do.||\n\"Fmaj7\" z2 A2 B2 c2 | d2 c2 A2 F2 |\"Bdim7\" E8 |\"E7\" B8 |\"Em7\" z2 E2 G2 A2 |\"A7\" B2 A2 G2 E2 | %22\nw: I'm glad that|I'm the one who|found|you,|that's why I'm|al- ways han- gin'|\n\"D7\" D8 |\"Dm7\"\"G7\" G8 |\"Cmaj7\" A c3- c4 |\"Gm7\" A G3-\"C7\" G4 |\"Fmaj7\" E D3- D4 |\"Fm\" E D3- D4 | %28\nw: 'round|you.|Do I *|love you? *|Oh my, *|do I? *|\n\"Em7\" E G3-\"A7\" G4 |\"Dm7\" C4\"G7\" D4 |\"C\" C8 |] %31\nw: Hon- ey *|'deed I|do.|\n\n"
}
```


Figure 7: Metadata examples from the M4-RAG and WikiMT-X datasets. The top section shows an entry for "Mairi's Wedding" from the M4-RAG dataset, including detailed multilingual metadata in English and Vietnamese, and an associated audio recording identified by a YouTube ID. The bottom section presents an entry for "Deed I Do" from the WikiMT-X dataset, which includes a YouTube ID linking to an audio recording, a genre label (Jazz, one of eight predefined categories), four types of long-form text annotations, and a lead sheet in ABC notation.

Table 5: Average cosine similarity between text and music features across datasets and annotation types, alongside human ratings of music-text alignment and musical aesthetics. Cosine similarity reflects pairing quality as estimated by text-to-music retrieval models, while human ratings assess how well the text semantically aligns with the music (**Alignment**) and the perceived aesthetic quality of the music (**Aesthetics**).

Dataset	Annotation	CLaMP	CLaMP 2	CLaMP 3^{c2}_{sa}	CLaMP 3_{saas}	Alignment	Aesthetics
WikiMT	*Caption*	0.1900	0.1244	0.2028	0.2184	**4.83**	3.42
MidiCaps	*Caption*	0.1133	0.0255	0.1401	0.1583	3.92	2.83
WikiMT-X	*Background*	**0.2429**	0.1343	0.2239	0.2264	4.67	**3.50**
	Analysis	0.1336	0.1097	0.2261	0.2461	3.75	
	Description	0.0794	0.0451	**0.2359**	0.2752	3.42	
	Scene	0.0779	**0.1410**	0.2240	**0.2874**	3.67	
Dataset	**Annotation**	**CLAP**	**TTMR++**	**CLaMP 3^{c2}_{sa}**	**CLaMP 3_{saas}**	**Alignment**	**Aesthetics**
SDD	*Caption*	0.3446	0.3340	**0.1129**	0.1683	2.25	2.50
MusicCaps	*Caption*	0.2518	0.3957	0.1077	0.1500	4.08	2.58
WikiMT-X	*Background*	**0.3734**	**0.4477**	0.0092	0.1186	**4.50**	**3.58**
	Analysis	0.2594	0.3813	0.0024	0.1298	3.92	
	Description	0.2000	0.3340	0.0385	0.1738	3.33	
	Scene	0.1848	0.2590	0.0525	**0.1996**	3.83	

B Evaluation of Annotation Quality Across Benchmarks

To evaluate the semantic quality of music-text pairings, we conduct an evaluation combining (i) automatic similarity scores from models and (ii) human ratings on music-text alignment and musical aesthetics. The results are summarized in Table 5.

We include both LLM-generated and human-annotated datasets in this evaluation. Specifically, all symbolic datasets—WikiMT, MidiCaps, and WikiMT-X—contain text annotations generated by large language models. In contrast, the audio datasets differ in annotation quality: MusicCaps captions are written by trained musicians, while SDD was annotated by non-expert crowd workers.

For the automatic evaluation, we report the average cosine similarity between text and music embeddings generated by multiple retrieval models, including CLAP, TTMR++, CLaMP, CLaMP 2, and two CLaMP 3 variants. These similarity scores indicate how well the text and music are aligned in the shared embedding space.

WikiMT-X consistently achieves higher similarity scores across most annotation types compared to previous datasets. For example, its *Background* annotations reach cosine similarities of 0.2429 (CLaMP), 0.3734 (CLAP), and 0.4477 (TTMR++), outperforming scores seen in other datasets. Similar improvements are observed in the *Analysis* and *Description* categories, with the highest overall score of 0.2874 achieved by CLaMP 3_{saas} on the *Scene* annotation type.

To complement the automatic metrics, we conducted a human evaluation. Four conservatory-trained musicians (each with over 10 years of formal experience) rated 144 music-text examples based on two criteria: semantic alignment (how well the text matches the music) and musical aesthetics (the perceived quality of the music), using a 1-5 scale.

Results show that WikiMT-X performs on par with or better than expert-annotated datasets. In the symbolic domain, its *Background* annotations receive a high alignment score of 4.67—close to WikiMT's 4.83 and higher than MidiCaps' 3.92.

In the audio domain, WikiMT-X again demonstrates strong performance. Its alignment scores reach up to 4.50, and aesthetics up to 3.58—substantially outperforming SDD (2.25 / 2.50) and MusicCaps (4.08 / 2.58), which include more amateur or crowd-sourced material. The consistently higher aesthetic ratings for WikiMT and WikiMT-X likely stem from their inclusion of well-known Western popular music from the 20th century, in contrast to the less polished recordings found in other benchmarks.

Overall, the results support WikiMT-X as a high-quality benchmark for multimodal text-to-music retrieval. Its annotations show robust semantic alignment and musical relevance, confirmed by both model metrics and expert ratings. With careful filtering and validation, LLM-generated annotations can match or even exceed expert quality across both symbolic and audio datasets.

C *t*-SNE Visualizations on WikiMT-X

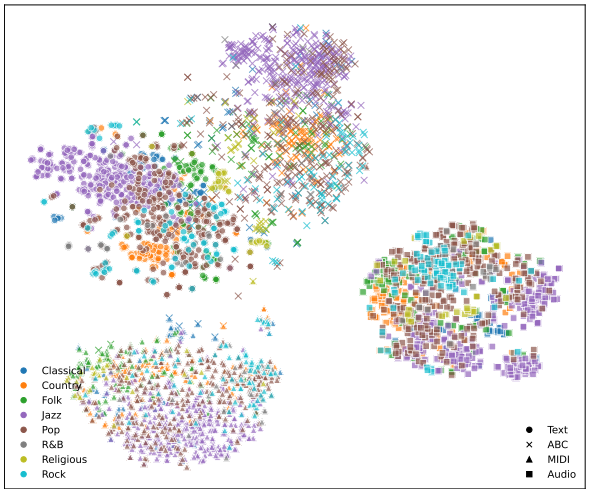
We apply *t*-SNE (t-distributed Stochastic Neighbor Embedding) to the WikiMT-X dataset to visualize how CLaMP 3 organizes data into a shared representation space. The projections illustrate the model's ability to align data across modalities, languages, and semantic categories.

Fig. 8a includes features from Text (background annotations), ABC notation, MIDI, and Audio. Each modality forms a distinct cluster, reflecting the inherent differences in how information is encoded. Notably, modalities closer to Text tend to perform better, aligning with the trend in Table 3, suggesting a correlation between embedding proximity and cross-modal effectiveness. Additionally, all musical modalities display a mirrored symmetry around the Text cluster, indicating that Text may serve as a semantic anchor. This symmetry suggests CLaMP 3 aligns modalities relative to Text, balancing modality-specific features while preserving semantic consistency.

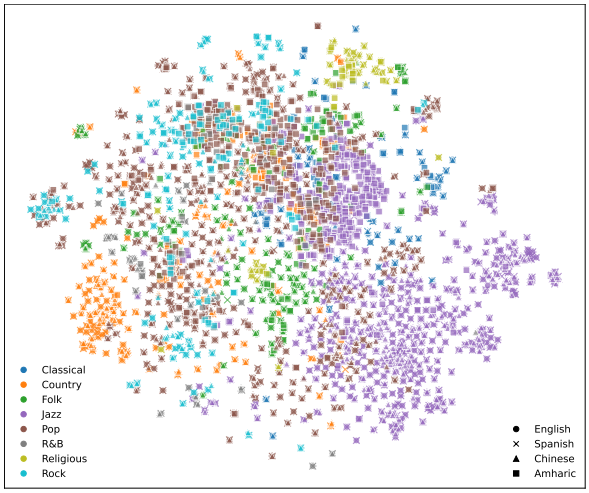
Fig. 8b focuses on background annotations in four languages—English, Spanish, Chinese, and Amharic—selected to represent varying retrieval performance levels. Despite their linguistic differences, these languages largely overlap, indicating strong cross-lingual alignment. English and Spanish cluster closely, reflecting both their shared linguistic roots. Chinese shows moderate overlap with English, suggesting that CLaMP 3 effectively bridges typologically distant languages. However, Amharic, a low-resource and unseen language, forms more isolated clusters, indicating the challenges of aligning low-resource languages.

Fig. 8c shows four semantic categories—*Background*, *Analysis*, *Description*, and *Scene*—presenting how CLaMP 3 handles different content types. *Background*, *Analysis*, and *Description* often converge, reflecting the overlap in explanatory texts as they cover related musical concepts. In contrast, *Scene* forms distinct clusters, likely because it focuses on visual depictions, leading to more consistent semantic patterns tied to specific imagery rather than music.

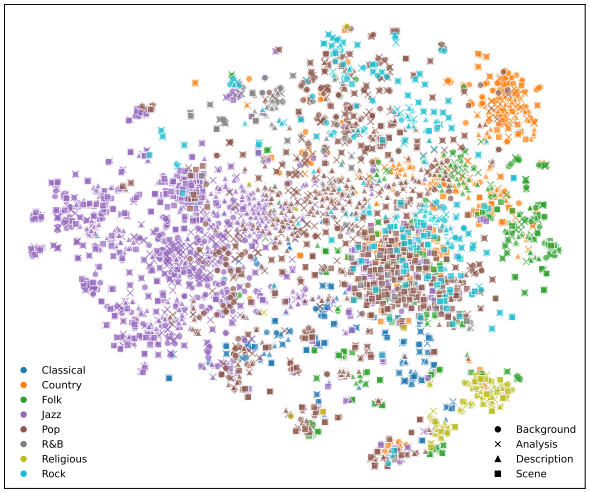
Across all three visualizations, genre boundaries remain clear despite differences in modality, language, or semantic category. This shows that CLaMP 3 effectively aligns multimodal and multilingual data while preserving genre-specific distinctions, demonstrating the model's strong representational capabilities.



(a) Modality



(b) Language



(c) Semantics

Figure 8: *t*-SNE visualization of the WikiMT-X dataset, illustrating the distribution of samples based on three distinct factors: (a) Modality, (b) Language, and (c) Semantics. The representations are extracted using CLaMP 3_{saas}. Each point represents a data sample, colored according to its genre.

Table 6: Results for English text-to-music retrieval on several benchmarks: WikiMT and MidiCaps have 1,010 pairs, Song Describer Dataset (SDD) has 706 audio and 1,106 captions, and MusicCaps-Remake (MC-R) contains 2,777 pairs. MC-R prevents data leakage by using full-length audio and rewritten captions from AudioSet's evaluation set.

Model	**Symbolic Benchmarks**		**WikiMT-X (Sheet Music)**			
	WikiMT	*MidiCaps*	*Background*	*Analysis*	*Description*	*Scene*
CLaMP 3_{as}	0.1973	0.0788	0.2108	0.1660	0.1049	0.1056
CLaMP 3_{sa}	0.3789	0.1322	0.3591	0.3088	0.1316	**0.1643**
CLaMP 3^{c2}_{sa}	**0.4498**	**0.2826**	**0.4028**	**0.3382**	0.0835	0.1512
CLaMP 3_{assa}	0.2993	0.0884	0.2919	0.2507	**0.1459**	0.1464
CLaMP 3_{saas}	0.3555	0.1798	0.3301	0.2758	0.1274	0.1512
CLaMP 3^{c2}_{saas}	0.3631	0.2688	0.3295	0.2957	0.0951	0.1395
Model	**Audio Benchmarks**		**WikiMT-X (Audio)**			
	SDD	*MC-R*	*Background*	*Analysis*	*Description*	*Scene*
CLaMP 3_{as}	0.1977	0.1117	0.1602	0.1375	0.0854	0.0819
CLaMP 3_{sa}	0.1607	0.0937	0.1718	0.1586	0.0997	0.0871
CLaMP 3^{c2}_{sa}	0.1612	0.0959	0.1180	0.1206	0.0639	0.0619
CLaMP 3_{assa}	0.2003	0.1045	0.1597	0.1522	**0.1020**	0.0873
CLaMP 3_{saas}	0.1985	0.1177	**0.2017**	**0.1711**	0.0988	**0.0963**
CLaMP 3^{c2}_{saas}	**0.2115**	**0.1180**	0.1583	0.1530	0.0768	0.0885

Table 7: Results for multilingual text-to-music retrieval on translated WikiMT-X background annotations. Languages marked with asterisks were not included in the M4-RAG training data. The BLEU scores below each language are calculated by back-translating the text with the SeamlessM4T model and comparing it to the original English text.

Model	**ru** *49.69*	**fr** *55.50*	**es** *62.82*	**ar** *53.38*	**zh** *39.58*	**fi*** *39.19*	**el*** *55.55*	**ta*** *40.07*	**kk*** *36.57*	**am*** *56.08*
ABC Notation										
CLaMP 3_{as}	0.1750	0.1931	0.1964	0.1594	0.1559	0.1828	0.1641	0.0997	0.1575	0.0876
CLaMP 3_{sa}	0.3262	0.3544	0.3536	0.3072	**0.2459**	0.3163	0.2879	0.1336	0.2894	0.1317
CLaMP 3^{c2}_{sa}	**0.3614**	**0.3949**	**0.3921**	**0.3155**	0.2373	**0.3524**	**0.3226**	0.1415	**0.3397**	**0.1871**
CLaMP 3_{assa}	0.2648	0.2810	0.2817	0.2450	0.2271	0.2644	0.2415	**0.1432**	0.2561	0.1300
CLaMP 3_{saas}	0.2918	0.3214	0.3239	0.2789	0.2358	0.2919	0.2681	0.1246	0.2703	0.1139
CLaMP 3^{c2}_{saas}	0.2954	0.3171	0.3225	0.2773	0.2144	0.2990	0.2721	0.1348	0.2750	0.1690
MIDI										
CLaMP 3_{as}	0.0418	0.0416	0.0432	0.0404	0.0332	0.0456	0.0449	0.0297	0.0398	0.0267
CLaMP 3_{sa}	0.1174	0.1284	0.1316	0.1132	0.0890	0.1217	0.1112	0.0623	0.1117	0.0540
CLaMP 3^{c2}_{sa}	**0.1921**	**0.2101**	**0.2137**	**0.1681**	**0.1316**	**0.2019**	**0.1702**	**0.0804**	**0.1765**	**0.1039**
CLaMP 3_{assa}	0.0565	0.0582	0.0620	0.0582	0.0517	0.0620	0.0585	0.0394	0.0595	0.0354
CLaMP 3_{saas}	0.1165	0.1319	0.1330	0.1141	0.0937	0.1245	0.1143	0.0601	0.1104	0.0544
CLaMP 3^{c2}_{saas}	0.1499	0.1645	0.1664	0.1408	0.1049	0.1560	0.1399	0.0653	0.1335	0.0793
Audio										
CLaMP 3_{as}	0.1267	0.1515	0.1525	0.1210	0.1089	0.1430	0.1428	0.0610	0.1043	0.0559
CLaMP 3_{sa}	0.1619	0.1717	0.1714	0.1529	0.1414	0.1585	0.1544	**0.0991**	0.1456	0.0774
CLaMP 3^{c2}_{sa}	0.1068	0.1150	0.1202	0.0981	0.0877	0.1112	0.1014	0.0720	0.1005	0.0681
CLaMP 3_{assa}	0.1426	0.1580	0.1588	0.1370	0.1202	0.1468	0.1431	0.0795	0.1276	0.0617
CLaMP 3_{saas}	**0.1788**	**0.1980**	**0.1962**	**0.1665**	**0.1459**	**0.1770**	**0.1736**	0.0945	**0.1561**	0.0675
CLaMP 3^{c2}_{saas}	0.1331	0.1566	0.1554	0.1304	0.1208	0.1550	0.1460	0.0901	0.1340	**0.0874**

Table 8: Results for emergent cross-modal retrieval on WikiMT-X pairings across different musical modalities. **S**: Sheet Music (ABC notation), **P**: Performance Signals (MIDI, converted from ABC), **A**: Audio recordings.

Model	S→P	S→A	P→S	P→A	A→S	A→P
CLaMP 3_{as}	0.1637	0.0557	0.1477	0.0248	0.0456	0.0237
CLaMP 3_{sa}	0.3205	**0.0739**	0.3054	0.0397	0.0479	0.0237
CLaMP 3^{c2}_{sa}	**0.4547**	0.0543	**0.5293**	0.0313	0.0492	0.0383
CLaMP 3_{assa}	0.1911	0.0619	0.1646	0.0299	0.0513	0.0264
CLaMP 3_{saas}	0.3262	0.0578	0.3146	0.0397	0.0410	0.0303
CLaMP 3^{c2}_{saas}	0.3909	0.0688	0.4375	**0.0467**	**0.0558**	**0.0431**

D Performance of CLaMP 3 Variants

A straightforward way to train CLaMP 3 would be to align symbolic music, audio, and text all at once. However, early experiments showed that this led to unstable training. The text encoder struggled because symbolic and audio data had very different distributions (Fig.8a) and pulled it in opposite directions, making alignment ineffective. To solve this, we adopted a multi-stage alignment strategy (Sec. 2.1) that gradually integrates each modality, ensuring stable and effective alignment.

To explore the best way to align modalities, we tested different training orders, leading to several model variants. The main difference among them is how and when the text encoder is aligned with symbolic music and audio encoders:

CLaMP 3_{as}: A two-stage alignment where text is first aligned with audio, then the text encoder is frozen while aligning with symbolic music.

CLaMP 3_{sa}: The reverse of CLaMP 3_{as}, first aligning text with symbolic music, then freezing the text encoder while aligning with audio.

CLaMP 3^{c2}_{sa}: Same as CLaMP 3_{sa}, but starting with pre-trained text and symbolic encoders from CLaMP 2.

CLaMP 3_{assa}: A four-stage alignment: audio → symbolic → symbolic → audio, with the text encoder frozen in the second and fourth stages to maintain stability.

CLaMP 3_{saas}: A four-stage alignment: symbolic → audio → audio → symbolic, also freezing the text encoder in the second and fourth stages.

CLaMP 3^{c2}_{saas}: Same as CLaMP 3_{saas}, but initialized with pre-trained text and symbolic encoders from CLaMP 2.

We evaluate these six variants across all experiments in Sec. 4 to assess their effectiveness in different retrieval tasks.

Table 6 shows that aligning text with symbolic music before audio improves generalization in English text-to-music retrieval. CLaMP 3_{sa} outperforms CLaMP 3_{as} in symbolic retrieval without compromising audio performance. Four-stage models outperform two-stage models in audio retrieval, emphasizing the importance of iterative alignment. Among them, CLaMP 3_{saas} achieves the best balance between symbolic and audio retrieval. Leveraging CLaMP 2's weight initialization enhances symbolic retrieval, as seen in CLaMP 3^{c2}_{sa} leading symbolic tasks. However, it does not consistently improve audio retrieval, likely because CLaMP 2 was trained only on symbolic music, limiting its text encoder's adaptability to audio alignment.

Table 7 demonstrates the impact of pre-training and training order on multilingual text-to-music retrieval. In symbolic retrieval, using CLaMP 2's pre-trained text-symbolic encoders provides a clear advantage, with CLaMP 3^{c2}_{sa} achieving the highest scores across most languages. This suggests that pre-training helps build a strong shared representation space, especially for MIDI, where M4-RAG's limited native data weakens overall performance. However, pre-training is not always decisive, as some non-pretrained models surpass pre-trained variants in certain languages for ABC retrieval. In contrast, audio retrieval is consistently strongest with CLaMP 3_{saas}, even in unseen languages, suggesting that training order plays a more crucial role in cross-lingual generalization.

Table 8 evaluates emergent cross-modal retrieval, where no direct supervised alignment exists among musical modalities. CLaMP 3^{c2}_{sa} achieves the best symbolic retrieval (S↔P), showing that CLaMP 2 pre-training strengthens symbolic-text alignment, which indirectly benefits symbolic retrieval. For symbolic-audio retrieval, CLaMP 3^{c2}_{saas} performs best, leading in P→A (0.0467), A→S (0.0558), and A→P (0.0431). It consistently outperforms CLaMP 3_{saas}, suggesting that pre-training provides a stronger shared representation space, leading to better cross-modal generalization between unpaired modalities.

These results show the importance of both training order and pre-training in MIR. Multi-stage alignment stabilizes training, while training order plays a key role, particularly in audio retrieval and cross-lingual generalization. Pre-training with CLaMP 2 strengthens symbolic retrieval and improves cross-modal generalization, but its benefits are limited for audio retrieval.

Table 9: Symbolic classification performance for ABC notation and MIDI was assessed across three datasets: WikiMT (1,010 pieces, 8 genres), VGMIDI (204 pieces, 4 emotions), and Pianist8 (411 pieces, 8 composers).

Model	Modality	WikiMT		VGMIDI		Pianist8	
		F1-macro	*Accuracy*	*F1-macro*	*Accuracy*	*F1-macro*	*Accuracy*
M3	ABC	0.2349	0.4010	0.6016	0.6341	0.7395	0.7590
CLaMP	ABC	0.3452	0.4267	0.6453	0.6866	0.7067	0.7152
CLaMP 2	ABC	**0.3990**	**0.4653**	0.7449	**0.8049**	**0.8025**	**0.8072**
CLaMP 3_{as}	ABC	0.3135	0.4307	0.6638	0.7073	0.6872	0.6867
CLaMP 3_{sa}	ABC	0.3225	0.4455	0.7725	**0.8049**	0.7403	0.7590
CLaMP 3^{c2}_{sa}	ABC	0.3316	0.4356	0.6845	0.7317	0.7722	0.7711
CLaMP 3_{assa}	ABC	0.3102	0.4455	0.4990	0.6341	0.6796	0.6988
CLaMP 3_{saas}	ABC	0.3177	0.4356	**0.7969**	**0.8049**	0.7716	0.7952
CLaMP 3^{c2}_{saas}	ABC	0.3568	0.4257	0.6694	0.7561	0.7891	0.7952
M3	MIDI	0.2621	0.4257	0.5399	0.6098	**0.9199**	**0.9157**
CLaMP 2	MIDI	0.2898	0.4455	0.5246	0.6585	0.8927	0.8916
CLaMP 3_{as}	MIDI	**0.3361**	**0.4653**	0.5600	0.5854	0.8186	0.8313
CLaMP 3_{sa}	MIDI	0.2614	0.4010	**0.6864**	**0.7073**	0.8461	0.8554
CLaMP 3^{c2}_{sa}	MIDI	0.3073	0.4455	0.6223	**0.7073**	0.8696	0.8675
CLaMP 3_{assa}	MIDI	0.2882	0.4406	0.5001	0.6098	0.8076	0.8193
CLaMP 3_{saas}	MIDI	0.2721	0.4158	0.5723	0.6341	0.7834	0.7952
CLaMP 3^{c2}_{saas}	MIDI	0.2943	0.4208	0.5474	0.6829	0.8565	0.8554

Table 10: Audio classification performance is evaluated on multiple benchmarks included in MARBLE: MTT (25,860 clips, 50 tags), GS (7,035 clips, 24 keys), GTZAN (1,000 clips, 10 genres), EMO (744 clips, valence/arousal regression), Nsynth (305,979 clips, 11 instrument categories, 88 pitches), and VocalSet (7,506 clips, 17 singing techniques, 20 singers).

Model	MTT Tagging		GS Key	GTZAN Genre	EMO Emotion		Nsynth Instrument	Nsynth Pitch	VocalSet Tech	VocalSet Singer
	ROC	*AP*	*Acc*	*Acc*	$R2^V$	$R2^A$	*Acc*	*Acc*	*Acc*	*Acc*
$MERT_{mean}$	0.9068	0.3915	**0.6475**	0.6689	0.5185	**0.7501**	0.6963	**0.9152**	**0.7219**	**0.8961**
CLAP	0.9066	0.3897	0.1596	0.8207	0.5408	0.7025	**0.7817**	0.5146	0.6868	0.6327
TTMR++	0.9082	0.3922	0.1672	0.8551	0.5599	0.7116	0.6735	0.5012	0.6342	0.5352
CLaMP 3_{as}	0.9097	0.3888	0.4935	0.8379	0.5944	0.7413	0.6445	0.8601	0.6780	0.8491
CLaMP 3_{sa}	0.9084	0.3863	0.2533	0.8448	**0.6031**	0.6949	0.6338	0.8647	0.7061	0.8419
CLaMP 3^{c2}_{sa}	0.9092	0.3924	0.2545	0.8551	0.5477	0.6876	0.6147	0.8574	0.6710	0.8007
CLaMP 3_{assa}	0.9098	0.3935	0.1498	**0.8793**	0.5921	0.7327	0.6411	0.8742	0.6842	0.8555
CLaMP 3_{saas}	**0.9109**	**0.3941**	0.5377	0.8655	0.5907	0.7004	0.6377	0.8689	0.7053	0.8441
CLaMP 3^{c2}_{saas}	0.9095	0.3938	0.3907	0.8138	0.5368	0.6589	0.6562	0.8732	0.6798	0.8470

Table 11: Audio classification performance on the MTG-Jamendo dataset (55,000+ tracks) was evaluated across four tasks: instrument classification (41 tags), mood/theme classification (59 tags), genre classification (95 tags), and top-50 multi-label classification.

Model	Instrument		Mood/Theme		Genre		Top50	
	ROC	*AP*	*ROC*	*AP*	*ROC*	*AP*	*ROC*	*AP*
$MERT_{mean}$	0.7421	0.1764	0.7598	0.1383	0.8672	0.1818	0.8280	0.2837
CLAP	0.7480	0.1812	0.7601	0.1323	0.8544	0.1716	0.8197	0.2773
TTMR++	0.7806	0.2111	0.7705	0.1477	0.8742	0.2030	**0.8340**	0.3049
CLaMP 3_{as}	0.7895	0.2254	0.7814	0.1476	0.8750	**0.2114**	0.8321	0.3068
CLaMP 3_{sa}	0.7780	0.2112	0.7823	0.1533	0.8713	0.2008	0.8276	0.3011
CLaMP 3^{c2}_{sa}	0.7832	0.2168	0.7796	0.1475	0.8679	0.2046	0.8220	0.2964
CLaMP 3_{assa}	**0.7911**	**0.2269**	0.7828	0.1486	**0.8763**	0.2109	0.8290	0.3041
CLaMP 3_{saas}	0.7872	0.2208	**0.7835**	**0.1547**	0.8703	0.2076	0.8242	0.3021
CLaMP 3^{c2}_{saas}	0.7803	0.2145	0.7825	0.1522	0.8734	0.2092	0.8296	**0.3074**

E Music Classification

This section evaluates CLaMP 3 variants and baselines via linear probing, assessing their ability to classify musical attributes in symbolic and audio music, as well as musical modalities and text annotations in WikiMT-X.

E.1 Symbolic Music Classification

Table 9 presents symbolic music classification results for ABC notation and MIDI across three benchmarks:

WikiMT (Wu et al., 2023a) consists of 1,010 lead sheets in ABC notation sourced from Wikifonia[4], labeled into 8 genre categories based on corresponding Wikipedia entries.

VGMIDI (Ferreira and Whitehead, 2019) contains 204 MIDI transcriptions of video game soundtracks, annotated with 4 emotion labels derived from valence and arousal levels.

Pianist8 (Chou et al., 2021) includes 411 piano performances, transcribed from audio to performance MIDI, and labeled with their respective composers across eight categories.

To enable evaluation in both formats, all datasets were converted between ABC and MIDI.

Despite improved text alignment, CLaMP 3 does not surpass CLaMP 2 in sheet music classification. This is likely because CLaMP 3 was trained on only half as much symbolic data. While stronger textual supervision benefits retrieval, it does not fully offset the reduced symbolic training for classification. However, CLaMP 3 still outperforms M3—the symbolic music encoder it was initialized from—on most benchmarks, suggesting that contrastive text supervision enhances the semantic salience of extracted features.

These results indicate that retrieval and classification improvements are relatively independent. In text-to-music retrieval (Table 2, Table 3), CLaMP 3—especially CLaMP 3^{c2}_{sa}—significantly outperforms CLaMP 2, yet this advantage does not extend to classification. A possible explanation is that retrieval requires rich representations and effective interaction between text and music encoders, while classification depends solely on an encoder's ability to extract features relevant to predefined labels. Thus, while higher-quality text annotations enhance retrieval, they do not necessarily improve symbolic music classification.

[4] http://www.synthzone.com/files/Wikifonia/Wikifonia.zip

E.2 Audio Music Classification

To evaluate the audio classification performance of CLaMP 3 variants and baselines, we conduct linear probing on MARBLE (Yuan et al., 2023) and MTG-Jamendo (Bogdanov et al., 2019).

MARBLE is a comprehensive benchmark collection for music representation evaluation. We assess models on 8 tasks covering different aspects of audio understanding. MTG-Jamendo is a large-scale benchmark with over 55,000 music tracks annotated for multiple classification tasks. It focuses on high-level musical attributes, making it well-suited for evaluating a model's ability to capture semantic meaning in music.

We also assess the self-supervised model MERT, CLaMP 3's audio feature extractor, averaging embeddings to one per 5-second clip across layers and time steps.

Table 10 shows the strengths of contrastive and self-supervised models on the MARBLE benchmark. CLaMP 3 variants excel in high-level tasks, like genre classification (GTZAN) and tagging (MTT), where capturing abstract musical meaning is crucial. MERT, however, performs better in low-level tasks such as key detection (GS) and pitch classification (Nsynth), where fine spectral detail is more important. Contrastive models generally struggle with short-duration audio (e.g., 4-second clips in Nsynth) because their focus on aligning longer segments with text limits their ability to capture acoustic details. These results suggest contrastive learning is better for semantic tasks, while self-supervised models are more effective for low-level acoustic analysis.

Table 11 shows that contrastive models, particularly CLaMP 3 variants, consistently outperform MERT across all MTG-Jamendo tasks. Notably, CLaMP 3 models achieve the highest scores in most tasks, demonstrating how diverse and high-quality text annotations help contrastive models learn and capture complex musical semantics.

In summary, contrastive models perform well in high-level classification tasks but struggle with short clips and fine-grained acoustic details. Their effectiveness heavily depends on the text annotations used during training. For instance, CLAP achieves strong results in instrument classification (Nsynth) because its training data is dominated by instrument and genre descriptions. However, it performs poorly in key detection (GS), where such annotations offer little relevant information.

Table 12: Classification performance on WikiMT-X (1,000 entries, 8 genres) across different musical modalities and text annotations.

Model	ABC	MIDI	Audio	Background	Analysis	Description	Scene
Accuracy							
CLaMP	0.7000	-	-	0.8050	0.7900	0.6900	0.6250
CLaMP 2	0.6800	0.6350	-	0.7900	0.8150	0.7250	0.6150
CLAP	-	-	0.6450	0.6950	0.6800	0.6500	0.5550
TTMR++	-	-	0.7150	0.7400	0.7600	0.6700	0.5950
CLaMP 3_{as}	0.6850	0.6100	0.7050	0.8200	0.8350	**0.7800**	0.6550
CLaMP 3_{sa}	0.7000	0.6650	0.6850	0.8000	0.8600	0.7700	0.6500
CLaMP 3^{c2}_{sa}	0.6850	0.6350	0.6850	0.7850	0.8550	0.7750	0.6500
CLaMP 3_{assa}	0.7000	0.6300	**0.7200**	**0.8650**	**0.8650**	0.7700	**0.6850**
CLaMP 3_{saas}	**0.7150**	**0.6800**	0.7050	0.8400	0.8550	**0.7800**	0.6650
CLaMP 3^{c2}_{saas}	0.6750	0.6300	0.6850	0.8300	0.8500	0.7700	**0.6850**
F1-macro							
CLaMP	0.5252	-	-	0.6835	0.6486	0.6079	0.4447
CLaMP 2	0.5287	0.3784	-	0.6617	0.6832	0.6333	0.3710
CLAP	-	-	0.3943	0.5913	0.5491	0.4921	0.3100
TTMR++	-	-	0.4714	0.6914	0.6694	0.6254	0.4246
CLaMP 3_{as}	0.5431	0.4005	0.4755	0.7424	0.7933	**0.7639**	0.4780
CLaMP 3_{sa}	0.5345	0.5108	0.4881	0.7917	0.8199	0.7372	0.4527
CLaMP 3^{c2}_{sa}	0.5428	0.4171	0.4589	0.6626	0.7439	0.7318	0.4260
CLaMP 3_{assa}	0.5499	0.3976	**0.5130**	**0.8486**	**0.8277**	0.6878	**0.5207**
CLaMP 3_{saas}	**0.5720**	**0.4967**	0.4995	0.8123	0.8225	0.7484	0.4742
CLaMP 3^{c2}_{saas}	0.5182	0.4313	0.4432	0.7811	0.8054	0.7082	0.4999

E.3 Classification on WikiMT-X

Table 12 presents classification results across different musical modalities (ABC, MIDI, Audio) and text annotations (*Background*, *Analysis*, *Description*, *Scene*) on WikiMT-X.

Compared to the WikiMT results in Table 9, all models show substantial gains in genre classification accuracy and F1-macro for ABC and MIDI. This confirms that reannotating genre labels significantly reduced label noise, leading to more reliable classification. The improvements suggest that earlier inconsistencies in genre annotations were a major limiting factor in classification performance. The reorganized label taxonomy and refined annotations in WikiMT-X provide a more structured and consistent genre framework, making it a more reliable benchmark for music classification.

Across different musical modalities, the best-performing models for ABC, MIDI, and Audio achieve comparable classification results. This suggests that genre-related features are well-preserved regardless of musical representation. Fig. 8a further supports this observation, showing clear genre boundaries across all modalities, indicating CLaMP 3 models can effectively extract genre information from both representations, reinforcing the idea that genre characteristics are consistently encoded in musical data.

A clear distinction emerges between text and music classification: models perform significantly better on text annotations (*Background*, *Analysis*, *Description*) than on music data. This is likely because text often contains explicit genre-related cues, making classification more direct. For example, descriptions like "syncopated piano chords and walking bass" strongly suggest jazz. In contrast, classifying music requires models to infer genre from intricate relationships between harmony, rhythm, and timbre. However, *Scene* classification behaves differently from other text-based categories—it describes environmental settings rather than musical attributes, making its classification challenge more similar to music than text.

Models trained solely on audio-text alignment (i.e., CLAP, TTMR++) perform worse in text classification, likely due to the limited diversity of annotations in large-scale audio-text datasets, which often list only instruments and genres. In contrast, symbolic-text datasets provide richer semantics, including background context and musicological analysis. CLaMP 3_{as} is an exception—though its text encoder was fully updated during audio alignment, it achieves much stronger text classification than models like CLAP and TTMR++. This is likely due to M4-RAG's well-curated and diverse annotations, which offer a broader and more expressive linguistic representation of musical content.

Table 13: Results for English text-to-music retrieval on MusicCaps, reflecting data leakage in baseline models. Evaluations are conducted on both the full set and the AudioSet evaluation set. R/O denotes the use of rewritten or original captions, while F/C indicates retrieval using full tracks or clips.

Model	Full Set (5,521 pairs)				Eval Set (2,858 pairs)			
	RF	*RC*	*OF*	*OC*	*RF*	*RC*	*OF*	*OC*
CLAP	0.0536	0.0743	0.0640	0.0894	0.0657	0.0886	0.0774	0.1113
TTMR++	**0.1410**	**0.2315**	**0.1757**	**0.3155**	**0.1248**	**0.1341**	**0.1219**	**0.1382**
CLaMP 3_{as}	0.0874	0.0642	0.0696	0.0536	0.1119	0.0830	0.0917	0.0699
CLaMP 3_{sa}	0.0741	0.0591	0.0530	0.0431	0.0934	0.0735	0.0661	0.0572
CLaMP 3^{c2}_{sa}	0.0729	0.0609	0.0619	0.0504	0.0961	0.0832	0.0822	0.0651
CLaMP 3_{assa}	0.0830	0.0592	0.0743	0.0530	0.1045	0.0784	0.0897	0.0723
CLaMP 3_{saas}	0.0890	0.0705	0.0652	0.0523	0.1177	0.0889	0.0890	0.0682
CLaMP 3^{c2}_{saas}	0.0973	0.0737	0.0762	0.0550	0.1180	0.0933	0.0961	0.0710

F Data Leakage of MusicCaps

MusicCaps, a widely used text-to-music retrieval benchmark, includes 5,521 music-text pairs with 10-second audio clips. As a subset of AudioSet, many models are trained on overlapping data, raising concerns about reliability, as they may memorize seen examples rather than learning true retrieval patterns.

Table 13 shows text-to-music retrieval results on MusicCaps, examining data leakage in baseline models. We evaluate performance on the full dataset (Full Set) and the AudioSet evaluation subset (Eval Set), while also assessing the effects of caption rewording (Original vs. Rewritten) and audio length (Clip vs. Full Track).

Leakage varies across models: TTMR++ is the most affected, having been trained on MusicCaps pairs from the training set of AudioSet, exposing it to half the benchmark; CLAP, trained on the full AudioSet, has seen all MusicCaps audio; in contrast, CLaMP 3 has minimal exposure, with only 150 audio recordings appearing in M4-RAG.

To mitigate leakage effects, we introduce rewritten captions generated using Qwen, ensuring semantic consistency while incorporating structured aspect lists—detailed annotations of key musical attributes such as instrumentation, mood, and rhythm. Additionally, we conduct retrieval on both 10-second clips and full-length tracks, forming four evaluation settings:

- **RF:** Rewritten captions with full tracks.
- **RC:** Rewritten captions with clips.
- **OF:** Original captions with full tracks.
- **OC:** Original captions with clips.

Table 13 reveals clear data leakage. TTMR++ is the only model that performs worse on the evaluation set than on the full benchmark, despite the evaluation set containing fewer retrieval candidates, which should naturally lead to higher MRR scores. This suggests severe overfitting to seen MusicCaps training data. Additionally, both TTMR++ and CLAP show performance drops with rewritten captions and full-length tracks. For TTMR++, this suggests that these modifications help reduce leakage effects, though not entirely. For CLAP, the decline is likely due to rewritten captions incorporating more detailed semantic information from aspect lists, which may shift retrieval behavior.

In contrast, all CLaMP 3 variants show improved performance with rewritten captions, likely due to M4-RAG's use of Qwen, making them more attuned to its text patterns. They also gain an advantage in full-track retrieval. While baseline models rely on 10-second clips and average embeddings across segments, CLaMP 3 processes up to 640 seconds of audio, enabling it to capture relationships across an entire track. In contrast, baselines extract semantics from isolated clips, restricting their ability to utilize long-form audio context effectively.

These results raise broader concerns about benchmark reliability in text-to-music retrieval. Other benchmarks also face leakage risks—SDD, for instance, comes from MTG-Jamendo, which was included in CLAP's training data. In contrast, WikiMT-X, manually curated for this study, mitigates leakage by sourcing audio from the web rather than existing datasets. However, since this audio remains publicly accessible, large-scale models may still have exposure. To further reduce leakage, future benchmarks should prioritize private or newly recorded datasets for unbiased evaluation.