Lyrics Are Meant to Be Sung: Modeling Singable Similarity for Cover Song Identification with Lyrics

Anonymous ACL submission

Abstract

Lyric similarity is commonly modeled through various NLP models, implicitly treating lyrics as ordinary text. However, lyrics are written to be sung, and their similarity is fundamentally constrained by singability: whether different sentences can be performed over the same melody. We reconceptualize lyric similarity as singable similarity and propose a framework that learns lyric representations grounded in this principle. Our approach leverages lyric translation pairs that are optimized to fit the same melody, providing natural supervision for learning sentence-level similarity under singability constraints. We introduce a singability encoder model that jointly captures semantic content and syllable structure. Evaluated on line-level translated lyric retrieval and lyric-based cover song identification (CSI), our singability-aware representations consistently perform well on lyric tasks. These results highlight singability as a crucial dimension for lyric-based music information retrieval.

1 Introduction

In much of the recent research, lyrics are commonly processed as ordinary text in music information retrieval, and lyric similarity is modeled using standard natural language processing techniques. However, lyrics are written to be sung, and their form and content are constrained by a fixed melody. This melody imposes requirements on syllable count, stress pattern, rhythm, and rhyme; consequently, lyric similarity cannot be fully captured by textual overlap alone. As a result, treating lyrics as ordinary text is insufficient. To reflect how lyrics are actually used in songs, similarity should be based on "singability" features rather than textual similarity alone.

This constraint becomes particularly evident when the same song is performed in different languages. While the melody remains unchanged, the lyrics must be adapted to fit the musical structure. Direct translation prioritizes semantic fidelity, but singable lyrics often require systematic deviations from literal meaning to satisfy melodic constraints. As a result, two lyric lines may differ substantially at the textual level yet remain functionally equivalent when sung over the same melody. Recent work has introduced multilingual datasets of singable lyric translations across languages, including large-scale K-pop lyric translation datasets, enabling empirical study of this phenomenon.

This phenomenon can be illustrated using the K-pop lyric translation dataset (Kim et al., 2024), which consists of official Korean and English lyrics for songs released in both languages. We use the song "Cry for Me" by TWICE as an example. The original Korean lyric line, "너의 곁에 있어줄게" (/neo-ui gyeo-te i-sseo jul-gae/), consists of eight syllables and is tightly constrained by the melody. When adapted into English for performance, it is rendered as "Imma pretend we're going strong," which preserves the same syllable count and aligns with the melodic stress pattern, despite not being a literal translation of the original line. A more semantically faithful translation of the Korean lyric would be "I'll stay by your side." However, this version fails to satisfy the melodic constraints of the song, leading to mismatched syllable timing and stress, and is therefore unsingable under the given melody. This example illustrates that equivalence under melodic constraints is not determined by semantic similarity alone: lyrics that are not literal translations may be preferable for performance, whereas semantically closer translations can be unsuitable when singability is taken into account.

Existing approaches to lyric-based music retrieval largely overlook this distinction. Methods based on semantic text similarity (Fell and Sporleder, 2014) treat lyrics as ordinary text and cannot distinguish between semantically similar lines that differ in their functional suitability for be-

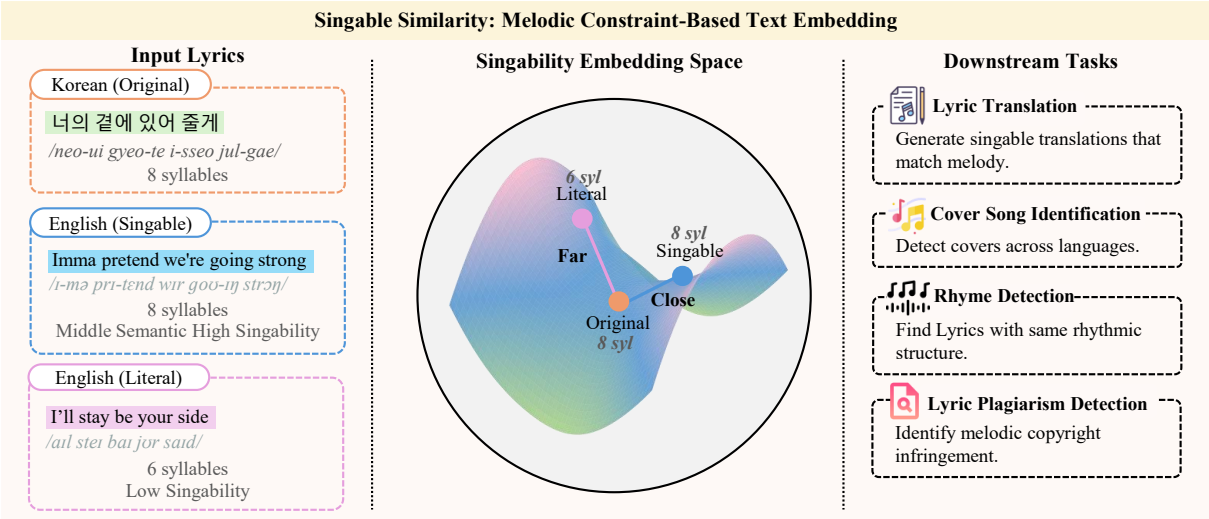

Figure 1: **Illustration of singable similarity.** In the learned embedding space, lyric lines that can be sung to the same melody are positioned close together, regardless of language or literal meaning.

ing sung. Conversely, acoustic approaches to tasks such as cover song identification (CSI) (Serra et al., 2008) operate purely on audio features, ignoring the structural and linguistic information encoded in lyrics. As a result, neither line of work adequately captures similarity grounded in musical performance.

Singable Similarity. The importance of fitting lyrics to melody, often referred to as "*singability*", has long been acknowledged in lyric translation and songwriting research (Low, 2005; Franzon, 2008). These studies describe inherent trade-offs between semantic fidelity and musical fit, noting that successful lyric adaptation often prioritizes rhythmic and phonetic compatibility over literal meaning.

Recent empirical work has further quantified these observations. In particular, Kim et al. (Kim et al., 2024) analyze large-scale lyric translation data and show that singable translations systematically sacrifice line-level semantic accuracy while preserving syllable structure, stress patterns, and repetition. Together, this body of work suggests that similarity between lyrics should be defined not only by what they mean or how they sound in isolation, but by whether they can be sung over the same melody.

Motivated by this perspective, we conceptualize lyric similarity as "*singable similarity*": similarity defined by melodic compatibility under shared musical constraints. Singable similarity is not reducible to semantic similarity or syllable similarity alone; rather, it reflects a structured interaction between meaning preservation and syllable adaptation.

Figure 1 illustrates our notion of singable similarity. In the learned embedding space, lyric lines that can be performed over the same melody are positioned close together, even when they differ substantially in language or literal meaning. Conversely, lines that are semantically similar but rhythmically incompatible are mapped far apart. This property enables several downstream applications beyond cover song identification: by detecting whether two lyric lines share melodic compatibility, or identifying potential lyric rhyme plagiarism cases.

Our Approach. We formulate lyric-based CSI as a similarity learning problem under a fixed melody. Given songs that share the same melody but may differ in language or wording, our goal is to identify lyric lines that are equivalent in terms of singability, namely lines that can be naturally sung over the same musical phrase.

To support this formulation, we operate at the line level to enable fine-grained, melody-aware alignment between cover versions, ensuring explainability and robustness to structural variations.

Based on this problem setting, we model singable similarity using a representation learning framework grounded in lyric translation. We introduce a dual-encoder architecture that captures both semantic meaning and singability-related structure. A multilingual semantic encoder represents the overall meaning of a lyric line, while a trainable syllable encoder models rhythm, and stress patterns derived from representations. Training on aligned

lyric translation pairs encourages the model to focus on melodic compatibility rather than surface-level textual similarity.

We evaluate our approach on three tasks of increasing complexity... While CSI serves as a song-level validation of our learned representations, our primary focus remains on the utility and interpretability of line-level matching. Across all tasks, incorporating singability consistently outperforms semantic-only baselines, highlighting it as a crucial dimension for lyric-based music information retrieval.

2 Related Work

2.1 Lyric Translation and Singability

Research on lyric translation emphasizes that song lyrics cannot be treated as ordinary text (Low, 2005; Franzon, 2008). Unlike prose, lyrics must conform to a fixed melody, imposing constraints on syllable count, rhythm, stress, and rhyme. Consequently, preserving literal meaning often leads to translations that are difficult to sing, while producing singable lyrics requires deliberate structural adaptation.

This tension was systematically articulated by Low (2005), who proposed the pentathlon approach, defining five competing criteria for song translation: singability, sense, naturalness, rhythm, and rhyme. Subsequent work by Franzon (2008) analyzed how translators prioritize different criteria depending on the translation setting, highlighting that singability frequently overrides strict semantic fidelity.

Motivated by these linguistic insights, recent computational studies have operationalized singability as an explicit modeling objective. Guo et al. (2022) showed that tonal constraints must be considered when translating lyrics into Mandarin. Ou et al. (2023) framed lyric translation as a constrained generation problem, incorporating length and rhyme controls to achieve near-perfect structural alignment across languages. Li et al. (2023) further demonstrated that jointly learning melody and lyric representations improves translation quality under musical constraints. More recently, Ye et al. (2024) proposed a two-stage training framework with learned reward models, showing substantial gains over naïve fine-tuning.

Beyond generation, Kim et al. (2023) introduced the first evaluation framework specifically designed for singable lyric translation, proposing metrics that capture syllable alignment, phoneme repetition, musical structure, and semantic preservation. Their follow-up work (Kim et al., 2024) provided large-scale empirical evidence from a Korean–English dataset, showing that singable translations prioritize structural and phonetic compatibility over line-level semantic similarity.

Most recently, Cho et al. (2025) released MAVL, a multilingual audio-video lyrics dataset spanning five languages. Their results demonstrate that incorporating multimodal cues significantly improves singability and contextual coherence, reinforcing the view that lyrics must be modeled in relation to musical structure.

In contrast to this line of work, which focuses on generating singable translations, we leverage lyric translation data as supervision for learning "*similarity*" representations. This enables similarity-based retrieval tasks such as cross-lingual lyric matching and lyric-based cover song identification (CSI).

2.2 Cover Song Identification

Cover song identification (CSI) aims to detect different renditions of the same musical composition despite variations in key, tempo, instrumentation, structure, and language.

Audio-Based Approaches. Most CSI research focuses exclusively on audio signals. Early methods utilized hand-crafted features, where chroma-based representations and alignment techniques handled key and tempo variations (Serra et al., 2008; Ellis and Poliner, 2007). Subsequent works addressed scalability through frequency-domain and timbral shape modeling (Ellis and Thierry, 2012; Tralie and Bendich, 2015).

The rise of deep learning transitioned the field toward learned audio embeddings. Metric learning was adopted for cover similarity (Doras and Peeters, 2019), while the Da-TACOS benchmark (Yesiler et al., 2019a) demonstrated that multi-task learning improves cross-dataset generalization. Recently, attention-based models like CoverHunter (Liu et al., 2023) have achieved state-of-the-art performance.

Modern approaches emphasize scalability and representation robustness. Recent studies have integrated pretrained ASR models (Du et al., 2024) or leveraged massive datasets like Discogs-VI (Araz et al., 2024) to train more expressive neural models.

Despite these advances, audio-only methods often neglect lyrical content—essential for capturing semantic and structural cues, particularly in cross-

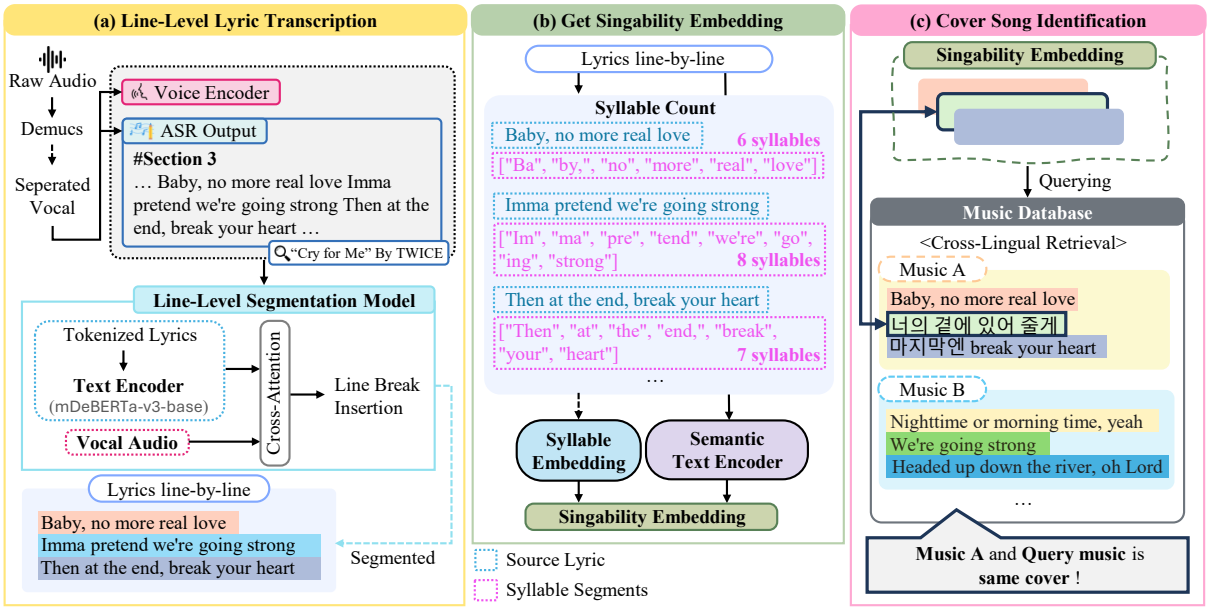

Figure 2: **Overview of the proposed singable similarity learning framework.** "Imma pretend we're going strong" is official cover lyric pair of "너의 곁에 있어 줄게".

lingual covers.

Lyric-Based Approaches. Vaglio et al. (2021) investigated lyric-based cover song detection, but their approach was limited to cases where cover versions share identical or nearly identical lyrics, and does not address scenarios involving lyrical adaptation or translation. This limitation becomes particularly critical in cross-lingual settings, where cover songs often preserve the melody while substantially modifying the lyrics to satisfy rhythmic, and linguistic constraints, making textual similarity an unreliable indicator of equivalence.

2.3 Cross-Lingual Text Representations

Multilingual language models have made significant progress in learning shared semantic spaces across languages. BERT and its multilingual variants demonstrated that cross-lingual transfer can emerge from shared pre-training objectives (Devlin et al., 2019; Conneau et al., 2020). Subsequent models such as LaBSE (Feng et al., 2022), SONAR (Duquenne et al., 2023), and E5 (Wang et al., 2022) further improved sentence-level semantic alignment using contrastive or translation-based objectives.

3 Problem Formulation

We study lyric-based cover song identification (CSI) in cross-lingual settings, where songs share the same underlying musical composition but differ in language and lyrical realization. Instead of generating translations, our objective is to retrieve cover songs by comparing lyrics at musically meaningful units. We decompose the problem into three tasks. (see Figure 2)

3.1 Line-Level Lyric Transcription

Input: An audio waveform $\mathbf{x}$.
Output: A sequence of line-level lyric segments $\{L_1, \ldots, L_m\}$.
Description: We first transcribe lyrics from the query music audio using an automatic speech recognition (ASR) system. Given a song and its lyrics, this task aligns the text to the audio and segments it into lyric lines. Unlike standard ASR, which produces word- or sentence-level transcripts, we target line-level units that correspond to musically meaningful phrases. These units typically align with melodic and rhythmic boundaries and serve as the fundamental granularity for downstream lyric retrieval.

3.2 Cross-Lingual Line-Level Lyric Matching

Input: A lyric line $L_i^{(A)}$ in language A and a lyric line $L_j^{(B)}$ in language B.

Output: A singable similarity score between $L_i^{(A)}$ and $L_j^{(B)}$.
Description: This task identifies singable similarity between two lyric lines. Matching must reflect singable similarity, such that two lines should be

compatible with the same melody flow. This requires modeling both semantic correspondence and syllable structure, such as syllable count and rhythmic pattern, which are not captured by semantic similarity alone.

3.3 Lyric-Based Cover Song Identification

Input: A query song Q and a database of candidate songs $\mathcal{D}$.
Output: A ranked list of cover song candidates from $\mathcal{D}$.
Description: Two songs are considered covers if they share the same underlying composition despite differences in arrangement, instrumentation, or language. Lyric-based CSI is performed by comparing the lyrics of Q against those in $\mathcal{D}$. Internally, line-level similarities obtained from cross-lingual lyric matching are aggregated into a song-level score while remaining robust to structural variations such as repetition, omission, or reordering of lyric lines.

4 Method

We present a detailed lyric-based framework for cross-lingual cover song identification (CSI) that operates on raw audio music.

4.1 Data Preprocessing

We begin by converting raw audio into lyrics if there's no lyric information. As in many audio–lyric-based studies (Vaglio et al., 2021), vocal tracks are extracted from polyphonic audio using htdemucs (Rouard et al., 2023) to reduce interference from accompaniment. For inference stage, when ground-truth lyrics are unavailable, we transcribe the full lyrics using Whisper large-v3 (Radford et al., 2023) with VAD filtering.

4.2 Line-Level Lyric Segmentation

Line-level lyric segmentation aims to recover musically meaningful lyric lines from audio, predicting where line breaks should occur based on both textual and acoustic cues. This task differs from conventional speech recognition, which focuses on word accuracy, and from music structure analysis, which operates at coarser section levels.

We formulate line segmentation as a simple word-level binary classification problem. Given a sequence of words, the model predicts for each word whether it is the last word of a line (label 1) or not (label 0). This formulation allows boundaries to be placed flexibly based on learned audio-text correspondences.

Model Architecture Our model fuses pretrained text representations with audio features through cross-modal attention. We employ multilingual-e5-large (Wang et al., 2024) as the text encoder, producing 1024-dimensional contextualized representations for each token. Audio features extracted from Whisper large-v3 encoder are projected to the same dimensionality via a linear layer. A multi-head cross-attention layer then allows text tokens to attend over the full audio sequence, enabling the model to capture acoustic cues such as pauses, pitch resets, and rhythmic boundaries. The attended features are combined with text representations through residual connection, and a linear classifier produces boundary predictions.

4.3 Cross-Lingual Line Representation Learning

Given segmented lyric lines, we learn representations that capture "*singable similarity*": whether two lyric lines can be sung to the same melody, regardless of language or lexical meaning.

We leverage aligned Korean–English lyric translation pairs as positive supervision. Unlike standard parallel text, these translations are explicitly constrained by melody, implicitly encoding syllable length, stress patterns, and rhythmic alignment.

Based on this supervision, we first derive line representations from a semantic feature extractor initialized with a pretrained multilingual language model, which captures meaning-level compatibility across languages. Token-level semantic representations are aggregated to form a line-level embedding, while the encoder is kept frozen during training to preserve general linguistic knowledge and avoid overfitting to limited lyric supervision.

Syllabic information is incorporated as a deliberately simple yet effective auxiliary cue, following recent research in lyric translation and representation learning that adopts syllable structure as a primary constraint (Kim et al., 2024; Cho et al., 2025). While we acknowledge that syllable count is a coarse proxy for melodic compatibility, it serves as a reasonable and efficient starting point that provides a fundamental constraint on melodic plausibility. This simple signal allows the model to down-weight lyric lines with highly implausible syllabic structures without the instability often associated with more complex phonetic features.

The semantic and syllabic representations are then concatenated to form the final line embedding. This design keeps semantic information as the pri-

Dataset	Lang.	Songs	Audio	Lyrics	Task Usage		
					LT	LM	CSI
K-pop Lyric Translation	KO–EN	1,000	Collected	✓	Train/Eval	Train/Eval	Train/Eval
JAM-ALT	Multi	79	✓	✓	Eval	–	–
Da-TACOS	EN	11,348†	Collected	×	–	–	Eval
Covers80	EN	160	✓	×	–	–	Eval

Table 1: Summary of datasets and their usage. LT = Line-level Transcription, LM = Cross-lingual Lyric Matching, CSI = Cover Song Identification. †Original dataset contains 15,000 songs; we use the 11,348 tracks with available audio.

mary representation while allowing simple syllabic cues to complement it, improving the model's ability to distinguish lines with different degrees of singability.

4.4 Song-Level Aggregation

For CSI, we aggregate line-level similarities into a song-level score using a Chamfer Distance-based metric. This approach accounts for the structural variations and line reorderings typical in cover versions. Given a query song Q with line embeddings $\{\mathbf{e}_i^Q\}_{i=1}^{|Q|}$ and a candidate song D with embeddings $\{\mathbf{e}_j^D\}_{j=1}^{|D|}$, the similarity score $S(Q, D)$ is defined as:

$$S(Q,D) = \frac{1}{2}\left(\frac{1}{|Q|}\sum_{i=1}^{|Q|}\max_j \cos(\mathbf{e}_i^Q, \mathbf{e}_j^D) + \frac{1}{|D|}\sum_{j=1}^{|D|}\max_i \cos(\mathbf{e}_i^Q, \mathbf{e}_j^D)\right) \quad (1)$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity. Unlike simple asymmetric pooling, this symmetric aggregation ensures that every line in both the query and the candidate is accounted for. It measures the mutual explanation between the two songs, making it highly robust to differences in repetition, structural shifts, and variations in lyric transcription across cover versions. Songs in the gallery are ranked in descending order of this score for retrieval-based evaluation.

5 Datasets

Table 1 summarizes the datasets and modalities used across lyric transcription, cross-lingual lyric matching, and CSI, all of which are split by song or musical work to prevent overlap between training and evaluation.

5.1 K-pop Lyric Translation Dataset

The K-pop Lyric Translation dataset (Kim et al., 2024) consists of 1,000 Korean songs paired with 1,000 singable English translations. Lyrics are aligned at both section and line levels, and 1,960 songs collected corresponding audio recordings. While the majority of songs are K-pop, the dataset also includes animated musical and theatrical songs.

This dataset supports all lyric-based tasks in this work, providing audio-lyric pairs for temporal segmentation in line-level transcription. For cross-lingual lyric matching, aligned Korean-English line pairs from the training split are used as positive examples for contrastive learning. For CSI, songs that share the same melody across languages are treated as cross-lingual cover song instances.

We split the dataset by song using an 8:1:1 ratio for training, development, and testing. Aligned lyric pairs from the development and test splits are used only for evaluation.

5.2 JAM-ALT

JAM-ALT (Cífka et al., 2024; Durand et al., 2023; Syed et al., 2025) is a readability-aware lyric transcription benchmark containing 79 songs in four languages (20 English, 20 German, 20 Spanish, and 19 French), with line-level timing annotations following industry transcription standards. We use JAM-ALT exclusively to define the evaluation protocol for line-level lyric transcription and do not train on this dataset.

5.3 Cover Song Identification Benchmarks

We evaluate CSI using two established benchmarks. Da-TACOS (Yesiler et al., 2019b) contains 15,000 songs. We used only 11,348 songs, which are available to download raw audio. Covers80 (Ellis, 2007) consists of 80 musical works with two versions each (160 songs in total). Both datasets are used only for evaluation. At inference time, the input

Dataset	Jam-ALT				K-pop trans.	
Method	WER↓	P_L↑	R_L↑	F_L↑	WER↓	F_L↑
Whisper v2	44.5	74.2	52.1	61.2	49.3	70.4
Whisper v3	48.0	76.3	57.6	65.7	49.4	70.4
AudioShake v3	**16.1**	**90.4**	**79.3**	**84.4**	–	–
Ours	30.9	88.0	59.4	71.0	**48.2**	**95.1**

Table 2: **Line-level segmentation results.** WER (%, ↓) and line break F-measure F_L (%, ↑). Bold indicates best performance per metric. Whisper v2/v3 (Radford et al., 2023) and AudioShake v3 (Cífka et al., 2024).

Method	mAP↑	R@1↑	R@5↑	R@10↑
Semantic-only	0.3486	0.2471	0.4742	0.5371
Singability	**0.3527**	**0.2520**	**0.4815**	**0.5402**

Table 3: **Cross-lingual line matching results** on the K-pop lyric translation (Kim et al., 2024) test set. We using the E5-large text encoder (Wang et al., 2024).

consists solely of a query song and a database of candidate songs; no lyric alignments or translation pairs are provided.

6 Experiments

We evaluate our framework on three tasks: (1) line-level lyric segmentation, (2) cross-lingual line matching, and (3) cover song identification (CSI). Each task targets a different component of singable similarity modeling, ranging from temporal lyric structuring to cross-lingual melodic equivalence and song-level retrieval.

6.1 Experimental Setup

Evaluation Metrics. For line-level segmentation, we compute precision, recall, and F1 with alt-eval, which gives WER and line-breaking evaluation. For cross-lingual line matching, we report mean Average Precision (mAP) and Recall@k. For CSI, we follow standard CSI evaluation protocols (**?**), reporting mAP and Mean Rank of the first correct result (MR1). All models were trained on a single NVIDIA RTX 5090 GPU.

6.2 Line-Level Segmentation

Setup. We trained a line segmentation model on the K-pop Lyric Translation dataset with collected raw audio. Audio inputs are segmented into fixed 30-second chunks, with up to 10 chunks used per song. The model is optimized using the AdamW optimizer with a learning rate of 2×10^{-5} and trained for up to 100 epochs. Training is performed with a token-level cross-entropy loss for line boundary prediction.

Baselines. We compare against representative lyrics transcription systems reported in the Jam-ALT benchmark, including Whisper v2/v3 (Radford et al., 2023), a general-purpose ASR model with demucs, and AudioShake v3 (Cífka et al., 2024) model. We also provide metrics with K-pop lyric test dataset.

Analysis. Table 2 reports line-level segmentation results on the JAM-ALT benchmark and the K-pop lyric test set. For our metrics, we focus on line-by-line transcription rather than full lyric transcription, which results in slightly lower overall scores. However, even when only raw audio is provided, the model is able to distinguish line boundaries effectively under a simple architecture. Despite being evaluated on unseen datasets and different languages, the performance remains competitive. We observe relatively high precision but lower recall in Jam-ALT (Cífka et al., 2024; Durand et al., 2023; Syed et al., 2025). We attribute this behavior to characteristics of K-pop music, where line segmentation occurs frequently, leading the model to produce conservative boundaries.

6.3 Cross-Lingual Line Matching

Setup. We evaluate cross-lingual line matching on the K-pop Lyric Translation test set, which contains aligned Korean–English lyric line pairs. Given a Korean query line, the task is to retrieve its corresponding English translation from a gallery consisting of all English lines in the test set. We using the AdamW optimizer with a batch size of 32. We apply a learning rate of 1×10^{-5} to the backbone encoder and 5×10^{-5} to the task-specific head. Models are trained for up to 30 epochs, and optimize a symmetric in-batch contrastive loss with a temperature of $\tau = 0.07$.

Baselines. We compare semantic-only multilingual text encoder E5-large (Wang et al., 2024) with their singability-aware counterparts. The singability models augment semantic representations with simple syllabic cues, while keeping the underlying text encoder unchanged.

Analysis. Table 3 presents cross-lingual line matching results on the K-pop lyric translation test set. Incorporating singability information leads to small but consistent improvements for E5-large

Method	Modality	Covers80		Da-TACOS(Test)		Da-TACOS-vocal		K-pop Lyrics	
		mAP↑	MR1↓	mAP↑	MR1↓	mAP↑	MR1↓	mAP↑	MR1↓
ByteCover2 (Du et al., 2022)	Audio	0.928	3.23	0.791	19.2	–	–	–	–
CoverHunter (Liu et al., 2023)	Audio	**0.933**	**3.20**	**0.865**	**11.0**	–	–	–	–
Vaglio et al. (Vaglio et al., 2021)	Audio+Lyrics	–	–	0.627	–	0.804	–	–	–
Ours	Lyrics	0.765	11.91	0.674	128.32	**0.896**	**25.04**	**0.993**	**1.08**

Table 4: **Cover song identification results across different benchmarks.** Best results for each column are highlighted in **bold**. We used groundtruth lyrics for K-pop lyrics dataset experiments, and Whisper-based lyric transcription for others.

across all retrieval metrics. This suggests that even minimal syllabic cues can help refine line-level matching beyond pure semantic similarity. Overall, these results show that singability-aware modeling can complement strong multilingual encoders. To further support these findings, we present a detailed result in Table 11.

6.4 Cover Song Identification

Setup. We evaluate cover song identification (CSI) on two standard benchmarks, Da-TACOS (Yesiler et al., 2019b) and Covers80 (Ellis, 2007). Additionally, we used ground-truth lyrics for the test K-pop lyric dataset (100 pair). For lyric-based CSI, lyrics are transcribed using Whisper and processed through our pipeline, which comprises line segmentation, extraction of singability embeddings with E5-large, and song-level aggregation. Since Da-TACOS provides metadata indicating whether a track is instrumental, we use the Da-TACOS-vocal subset, following the approach of Vaglio et al. (Vaglio et al., 2021).

Baselines. We compare against representative audio-based CSI methods, including ByteCover2 (Du et al., 2022) and CoverHunter (Liu et al., 2023), which learn robust audio embeddings for cover song retrieval. We also include the prior work explicitly incorporating lyrics into CSI, proposed by Vaglio et al. (Vaglio et al., 2021).

Analysis. Table 4 reports CSI results across multiple benchmarks. Our method differs in that it uses lyric representations designed to reflect whether lines can be sung to the same melody, rather than exact lexical overlap. In settings where reliable lyrics are provided, such as the Da-TACOS-vocal and K-pop Lyrics benchmarks, our approach achieves competitive or strong performance.

However, we observe a performance gap between the ground-truth (GT) K-pop dataset and the ASR-based benchmarks (Covers80 and Da-TACOS). To clarify the source of this degradation, we identify the upstream transcription quality as a critical bottleneck. Specifically, errors from Whisper-v3 and the subsequent line segmentation stage propagate to the similarity computation, particularly in the presence of background noise or diverse vocal styles. Furthermore, we cannot handle instrumental cases, which leads to major performance degradation on the full Da-TACOS test set. These results suggest that lyric-based CSI is most beneficial in scenarios where vocal content is present and lyrics can be obtained with sufficient accuracy, or by providing complementary cues alongside audio-based models to mitigate ASR-induced noise.

7 Conclusion

We argued that lyric similarity should not be treated as ordinary textual similarity, since lyrics are written to be sung under fixed melodic constraints. To address this gap, we introduced the concept of *singable similarity*, which captures whether different lyric lines can be performed over the same melody, beyond surface-level semantic overlap.

We proposed a singability modeling framework and its application, supervised by aligned singable lyric translations. This formulation enables line-level modeling that reflects melodic constraints.

Across line-level segmentation, cross-lingual line matching, and lyric-based cover song identification, our results show that incorporating singability leads to more meaningful lyric representations than semantic-only approaches, particularly in settings where reliable lyric information is available. These findings suggest that singability is a key dimension for lyric-based music information retrieval and should be explicitly modeled when lyrics are used as a primary signal.

8 Limitations

This work has several limitations that highlight important assumptions, scope constraints, and directions for future research.

Line-Level Transcription Scope and Upstream Assumptions. Our line-level lyric transcription task focuses exclusively on predicting line boundaries, rather than performing full lyric transcription directly from audio. The model assumes access to word-level lyric transcripts and does not jointly optimize speech recognition and segmentation. This assumption simplifies the problem setting and allows us to isolate structural modeling at the line level, but limits applicability in fully lyric-agnostic or end-to-end scenarios.

In practice, this assumption may be violated when lyric transcripts are unavailable or noisy. When lyrics must be obtained via automatic speech recognition, errors from upstream ASR systems (e.g., Whisper v3) can propagate to downstream segmentation, similarity computation, and cover song identification. Our results are therefore not robust to severe transcription errors, and performance degrades as ASR quality decreases, identifying upstream transcription as a primary bottleneck. Developing end-to-end frameworks that explicitly model uncertainty and jointly optimize transcription and musical structure remains an important direction for future work.

Simplified Singability Modeling. Our singability encoder explicitly models syllable counts as a primary constraint but abstracts away richer phonetic and prosodic phenomena such as rhyme, vowel quality, stress patterns, and consonant repetition. While syllable structure captures a core aspect of singability, it represents a simplified approximation of lyrical fit to melody.

We experimented with incorporating additional phonetic representations based on PWEsuite features, but these approaches did not yield consistent improvements and were less stable during training. As a result, we opted for a deliberately constrained design. This choice implies that the model may fail to distinguish between lines that are syllabically compatible but differ substantially in phonetic or poetic quality. Future work should explore more expressive phonetic representations that improve robustness without sacrificing training stability.

Modeling Level and Evaluation Mismatch. Although our approach emphasizes line-by-line modeling for interpretability, evaluation is primarily conducted on cover song identification (CSI), which is inherently a song-level retrieval task. This introduces a mismatch between the modeling granularity and the evaluation objective, requiring aggregation of line-level similarities into a single song-level score.

Our ablation studies show that models trained directly at the song level achieve higher CSI performance. This suggests that while line-level modeling may not be optimal for maximizing retrieval accuracy alone, it reflects a deliberate trade-off between absolute performance and explainability. We intentionally adopt this formulation to support fine-grained analysis and explainability, which are critical for downstream applications such as plagiarism detection, lyric reuse analysis, and editorial review. Nevertheless, the scope of our claims is limited to demonstrating the utility of line-level representations rather than establishing state-of-the-art CSI performance.

Dataset Constraints and Generalization. This study is constrained by the availability of datasets that jointly provide high-quality audio, aligned lyrics, and cross-lingual singable pairs. The K-pop lyric translation dataset offers reliable supervision for singable similarity but does not provide complete or uniform audio coverage. Other resources improve modality diversity but remain limited in scale or introduce variability through external audio acquisition, which may affect reproducibility.

As a result, our empirical findings are based on a limited set of datasets and languages and may not generalize uniformly to other musical genres, languages, or production styles. While some tasks, such as line-level segmentation, do not strictly require parallel lyric translations, broader validation on larger and more diverse audio–lyric corpora is necessary to assess robustness and general applicability.

Potential Risks and Misuse. Our work is intended as a research contribution toward understanding lyric structure, singability, and similarity, particularly for analysis and retrieval tasks. However, the techniques presented could potentially be misused in automated plagiarism accusation systems or content moderation pipelines if deployed without human oversight. Because our models operate under simplifying assumptions and may propagate upstream transcription errors, their outputs should not be interpreted as definitive judgments.

We emphasize that our approach is designed to support expert analysis rather than replace it. Future work should explore safeguards such as uncertainty estimation, human-in-the-loop workflows, and clearer usage guidelines to mitigate the risk of misinterpretation or overgeneralization in real-world settings.

References

R Oguz Araz, Xavier Serra, and Dmitry Bogdanov. 2024. Discogs-vi: A musical version identification dataset based on public editorial metadata.

Woohyun Cho, Youngmin Kim, Sunghyun Lee, and Youngjae Yu. 2025. Mavl: A multilingual audio-video lyrics dataset for animated song translation. pages 13651–13679.

Ondřej Cífka, Hendrik Schreiber, Luke Miner, and Fabian-Robert Stöter. 2024. Lyrics transcription for humans: A readability-aware benchmark.

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 ACL*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Guillaume Doras and Geoffroy Peeters. 2019. Cover detection using dominant melody embeddings. *arXiv preprint arXiv:1907.01824*.

Xingjian Du, Ke Chen, Zijie Wang, Bilei Zhu, and Zejun Ma. 2022. Bytecover2: Towards dimensionality reduction of latent embedding for efficient cover song identification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 616–620. IEEE.

Xingjian Du, Mingyu Liu, Pei Zou, Xia Liang, Zijie Wang, Huidong Liang, and Bilei Zhu. 2024. X-cover: Better music version identification system by integrating pretrained asr model. In *ISMIR*, pages 70–77.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Simon Durand, Daniel Stoller, and Sebastian Ewert. 2023. Contrastive learning-based audio to lyrics alignment for multiple languages. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Daniel P. W. Ellis. 2007. The "covers80" cover song data set. Web resource.

Daniel PW Ellis and Graham E Poliner. 2007. Identifyingcover songs' with chroma features and dynamic programming beat tracking. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–1429. IEEE.

Daniel PW Ellis and Bertin-Mahieux Thierry. 2012. Large-scale cover song recognition using the 2d fourier transform magnitude.

Michael Fell and Caroline Sporleder. 2014. Lyrics-based analysis and classification of music. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 620–631.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 878–891.

Johan Franzon. 2008. Choices in song translation: Singability in print, subtitles and sung performance. *The Translator*, 14(2):373–399.

Fenfei Guo, Chen Zhang, Zhirui Zhang, Qixin He, Kejun Zhang, Jun Xie, and Jordan Boyd-Graber. 2022. Automatic song translation for tonal languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 729–743.

Haven Kim, Jongmin Jung, Dasaem Jeong, and Juhan Nam. 2024. K-pop lyric translation: Dataset, analysis, and neural-modelling. *arXiv preprint arXiv:2309.11093*.

Haven Kim, Kento Watanabe, Masataka Goto, and Juhan Nam. 2023. A computational evaluation framework for singable lyric translation. *arXiv preprint arXiv:2308.13715*.

Chengxi Li, Kai Fan, Jiajun Bu, Boxing Chen, Zhongqiang Huang, and Zhi Yu. 2023. Translate the beauty in songs: Jointly learning to align melody and translate lyrics. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 27–39.

Feng Liu, Deyi Tuo, Yinan Xu, and Xintong Han. 2023. Coverhunter: Cover song identification with refined attention and alignments. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1080–1085. IEEE.

Peter Low. 2005. The pentathlon approach to translating songs. *Song and significance: Virtues and vices of vocal translation*, 25:185–212.

Longshen Ou, Xichu Ma, Min-Yen Kan, and Ye Wang. 2023. Songs across borders: Singable and controllable neural lyric translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–467.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Simon Rouard, Francisco Massa, and Alexandre Défossez. 2023. Hybrid transformers for music source separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Joan Serra, Emilia Gómez, Perfecto Herrera, and Xavier Serra. 2008. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1138–1151.

Jaza Syed, Ivan Meresman Higgs, Ondřej Cífka, and Mark Sandler. 2025. Exploiting music source separation for automatic lyrics transcription with whisper. In *2025 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE.

Christopher J Tralie and Paul Bendich. 2015. Cover song identification with timbral shape sequences.

Andrea Vaglio, Romain Hennequin, Manuel Moussallam, and Gael Richard. 2021. The words remain the same: Cover detection with lyrics transcription. In *22nd International Society for Music Information Retrieval Conference ISMIR 2021*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Zhuorui Ye, Jinhan Li, and Rongwu Xu. 2024. Sing it, narrate it: Quality musical lyrics translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5498–5520.

Furkan Yesiler, Chris Tralie, Albin Andrew Correya, Diego F Silva, Philip Tovstogan, Emilia Gómez Gutiérrez, and Xavier Serra. 2019a. Da-tacos: A dataset for cover song identification and understanding. In *Proceedings of the 20th Conference of the International Society for Music Information Retrieval (ISMIR 2019): 2019 Nov 4-8; Delft, The Netherlands.[Canada]: ISMIR; 2019*. International Society for Music Information Retrieval (ISMIR).

Furkan Yesiler, Chris Tralie, Albin Andrew Correya, Diego F Silva, Philip Tovstogan, Emilia Gómez Gutiérrez, and Xavier Serra. 2019b. Da-tacos: A dataset for cover song identification and understanding. In *Proceedings of the 20th Conference of the International Society for Music Information Retrieval (ISMIR 2019): 2019 Nov 4-8; Delft, The Netherlands.[Canada]: ISMIR; 2019*. International Society for Music Information Retrieval (ISMIR).

A Why Line-Level Segmentation?

Throughout this work, we model lyric similarity at the line level rather than aggregating entire songs into single embeddings. From a purely task-oriented perspective, song-level representations could achieve competitive performance on cover song identification (CSI). However, our primary motivation extends beyond retrieval accuracy. While identifying that song A' is a cover of song A is useful, explaining *why* they are covers requires interpretable representations that reveal *which parts* correspond and *how* they match despite differences in language or wording.

A.1 Explainability for Downstream Tasks

Line-level representations naturally extend to several tasks where local correspondence matters:

Lyric Plagiarism Detection: Identifying which specific lines exhibit suspicious overlap under melodic constraints, providing concrete evidence rather than song-level similarity scores.

Rhyme and Repetition Detection: Detecting rhyme schemes and phonetic patterns that operate at the line or sub-line level, useful for style analysis and generation.

Cross-Lingual Translation Evaluation: Providing interpretable feedback on which translated lines preserve singability and which deviate from melodic structure.

These applications require fine-grained alignment that song-level embeddings cannot provide. While this work focuses on cover song identification, the learned representations are designed to support these broader use cases.

A.2 Trade-offs

Line-level modeling increases computational cost from $O(1)$ to $O(|Q| \times |D|)$ comparisons per song pair and introduces sensitivity to segmentation errors. However, we believe interpretability benefits outweigh these costs for applications requiring transparency, such as copyright enforcement and musicological analysis.

A.3 Ablation Study

To rigorously analyze the impact of modeling granularity, we conducted an ablation study by training an additional song-level Siamese network as a performance upper-bound baseline. Unlike our main pipeline that aggregates line-level features, this model was specifically designed to learn holistic song representations. We concatenated the lyrics of each song into a single paragraph and fine-tuned a multilingual E5-large-instruct (Wang et al., 2024) backbone on the Da-TACOS (Yesiler et al., 2019b) training set. The training employed a Siamese architecture optimized via triplet loss, directly mapping entire lyric sequences into a shared latent space to maximize the similarity between cover-original pairs.

	Da-TACOS		Covers80	
Granularity	**mAP↑**	**MR1↓**	**mAP↑**	**MR1↓**
Song-level	**0.955**	**21.09**	**0.988**	**1.02**
Line-level	0.896	25.04	0.765	11.91

Table 5: Ablation study comparing song-level versus line-level granularity.

As shown in Table 5, our results indicate that the song-level model achieves the highest performance in cover song identification (CSI) tasks, significantly surpassing line-level aggregation. When a sufficient amount of lyric content is provided, this approach even exhibits to exceed current SOTA models. However, the song-level representation acts as a "black box", providing no information on which specific segments or phrases contribute to the similarity score. In contrast, while line-level modeling shows a slight trade-off in raw metrics, it enables precise, interpretable alignment between cover versions. Detailed score alignment in line-level modeling will be discussed in section F.

B Effects of vocal audio in Line-Level Segmentation

B.1 Observations

The visualization result in Figure 3 confirms the role of the attention mechanism in multimodal integration:

- **Acoustic Alignment:** Cross-modal attention peaks synchronize with the manually annotated segment boundaries, confirming that the audio signal provides effective temporal guidance.
- **Segment Impact:** The attention mechanism actively leverages acoustic cues to distinguish line-by-line transitions, facilitating precise alignment between the audio and textual modalities.

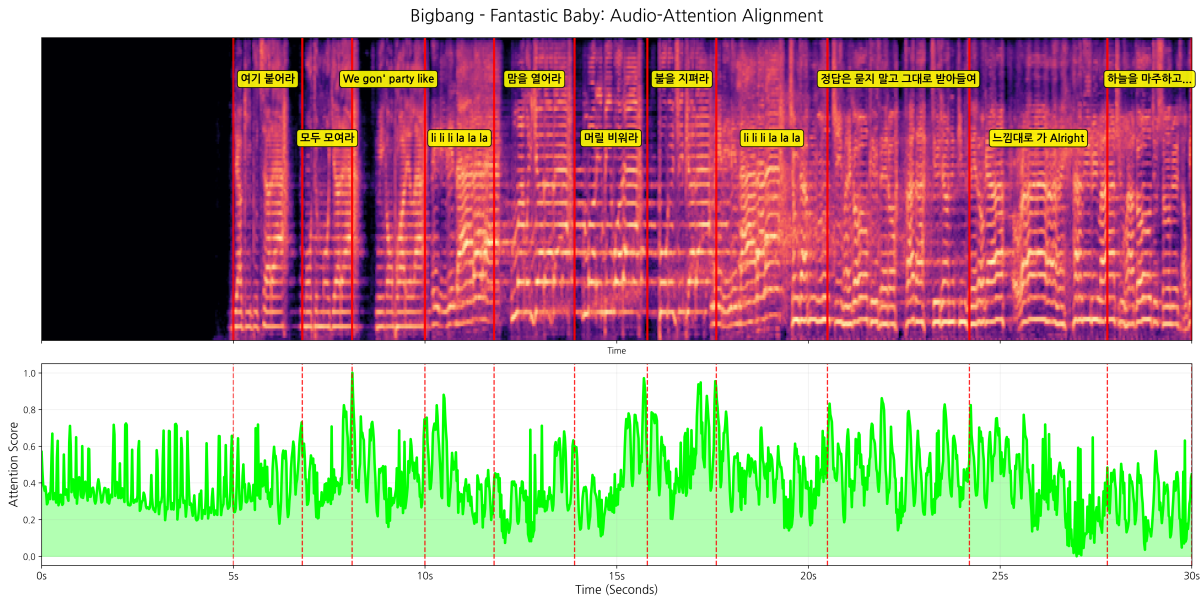

Figure 3: Visualization of cross-modal attention alignment. Red lines indicate line-by-line lyrical segment boundaries. Lyrical content is sourced from the Kpop Lyric Dataset, with timestamps manually annotated through auditory verification.

C Details in Line-level Segmentation

Raw audio is first processed using Demucs to separate the vocal track. The separated vocals are transcribed using a full-track ASR system; in our implementation, lyrics are extracted using the Whisper-large-v3 model (Radford et al., 2023). The resulting word sequence and vocal audio are then encoded by a text encoder and an audio encoder, respectively. These representations are fused through a multimodal feature fusion module to predict word-level line boundary labels. Finally, line breaks are inserted based on the predicted boundaries to produce line-segmented lyrics (see Figure 2(a)).

C.1 Additional Qualitative results

These results are obtained from YouTube-collected test data that are not included in the K-pop translation dataset (Kim et al., 2024) used for training and validation.

We present several examples consisting of original K-pop lyrics and their corresponding English versions. See Tables 7–10.

D Analysis of Syllable-Aware Semantic Attention

Our model is designed to preserve semantic similarity while accounting for syllabic constraints that affect singability. To this end, semantic representations are first extracted using a frozen multilingual language model, ensuring that meaning-level information is not distorted by lyric-specific supervision.

D.1 Effect of Syllable Count on Semantic Aggregation

Formally, given an input line x, the semantic encoder produces a sequence of token-level embeddings

$$\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N\}, \tag{2}$$

where $\mathbf{h}_i \in \mathbb{R}^d$ denotes the semantic embedding of the i-th token and N is the number of tokens in the line.

In parallel, we associate the line with a scalar syllable count $S \in \mathbb{N}$, which provides a simple cue about its syllabic structure. Importantly, S does not alter the semantic embeddings themselves; instead, it conditions how semantic information is aggregated.

The final line-level representation $\mathbf{v} \in \mathbb{R}^d$ is computed as an attention-weighted sum of token embeddings:

$$\mathbf{v} = \sum_{i=1}^{N} \alpha_i(S)\, \mathbf{h}_i, \tag{3}$$

where $\boldsymbol{\alpha}(S) = \{\alpha_1(S), \ldots, \alpha_N(S)\}$ is an attention distribution parameterized by the syllable count S and satisfies $\sum_i \alpha_i(S) = 1$.

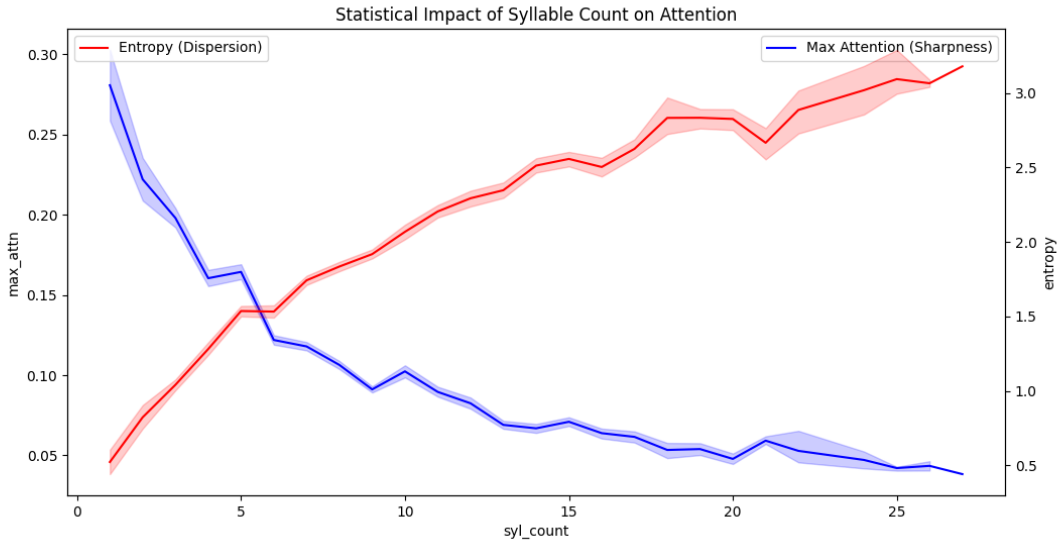

Figure 4: Statistical impact of syllable count S on attention mechanism. The blue line denotes Attention Sharpness ($\mathcal{S}$), while the red line denotes Attention Entropy ($\mathcal{H}$).

D.2 Syllable-Conditioned Attention Behavior

To examine how syllable information influences the attention mechanism, we analyze two statistics computed from the attention weights $\boldsymbol{\alpha}(S)$:

- **Attention Sharpness:** $\mathcal{S} = \max_i \alpha_i(S)$, indicating how strongly attention concentrates on a small number of tokens.
- **Attention Entropy:** $\mathcal{H} = -\sum_i \alpha_i(S) \log(\alpha_i(S) + \epsilon)$, measuring how broadly attention is distributed, where ϵ is a small constant for numerical stability.

As illustrated in Figure 4, smaller syllable counts result in higher sharpness and lower entropy, suggesting that the model focuses on a limited subset of semantic tokens.

In contrast, larger syllable counts lead to more dispersed attention, allowing semantic information to be aggregated over a wider temporal span. This trend indicates that syllable information modulates how semantic content is attended to, without explicitly encoding rhythmic rules.

D.3 Semantic Proximity vs. Singability-Aware Separation

An important consequence of this design is that semantic similarity and singability-aware distance are not necessarily aligned.

Semantically equivalent expressions such as "Hi" and "안녕하세요 (annyeonghaseyo)" are typically mapped to nearby regions by a standard semantic encoder, reflecting their shared meaning.

In our model, however, these expressions differ substantially in their syllable counts ($S = 1$ versus $S = 5$), which conditions the attention distribution used to aggregate semantic tokens. Let $\mathbf{v}_S$ denote the line-level representation obtained by syllable-conditioned attention:

$$\mathbf{v}_S = \sum_i \alpha_i(S)\, \mathbf{h}_i, \tag{4}$$

where $\mathbf{h}_i$ is the semantic embedding of the i-th token.

Because $\boldsymbol{\alpha}(1)$ and $\boldsymbol{\alpha}(5)$ exhibit markedly different sharpness and entropy characteristics, the resulting representations $\mathbf{v}_1$ and $\mathbf{v}_5$ become separated in the embedding space, despite their semantic proximity. This separation arises not from changes in semantic content, but from differences in how that content is selectively aggregated under syllabic constraints.

By conditioning attention on a single scalar syllable count, the model can effectively downweight lyric lines whose syllabic structure is implausible for a given melodic context, while still preserving semantic compatibility. As a result, the model acquires the ability to distinguish different degrees of singability using minimal phonetic information, without relying on hand-crafted rhythmic features

	K-pop Lyrics		Covers80		Da-TACOS (vocal)	
Method	mAP↑	MR1↓	mAP↑	MR1↓	mAP↑	MR1↓
Mean	0.9222	2.47	0.6078	14.65	0.8802	62.49
Max	0.8854	12.31	0.5360	21.18	0.8587	46.81
Top-5 Mean	0.6905	24.84	0.7419	13.07	0.8482	30.24
Hausdorff	0.8049	13.53	0.3305	22.38	0.6213	130.50
Soft-DTW	–	–	0.5008	17.61	–	–
Chamfer	**0.9925**	**1.08**	**0.7655**	**11.91**	**0.8956**	**25.04**

Table 6: Performance comparison of evaluated song-level aggregation methods. Soft-DTW values for large dataset cannot be computed due to computation resource.

or explicit alignment supervision.

E Detailed Analysis of Song-Level Aggregation Strategies

In this section, we provide a comprehensive comparative analysis of various song-level aggregation strategies. To justify the selection of the matching mechanism, we evaluate how different pooling and distance-based functions affect retrieval performance by fixing the underlying line-level encoder and varying only the aggregation logic.

E.1 Description of Evaluated Methods

We investigate five distinct strategies to translate a set of line embeddings into a final song-level similarity score:

- **Mean Pooling**: Computes the centroid of all line embeddings within a song to form a global representation.
- **Max Pooling**: Selects the maximum value across line embeddings for each dimension to capture the most salient features.
- **Top-5 Mean**: Calculates the average of the five highest similarity scores between any pair of lines from the two songs, focusing on strong local correspondences.
- **Hausdorff Distance**: Measures the maximum of the minimum distances between line sets, representing the worst-case misalignment between two lyrics.
- **Chamfer Distance**: A bidirectional aggregation that calculates the average of the maximum similarities between all lines of two songs, accounting for mutual explanation.

The performance across three different benchmarks is summarized in Table 6.

E.2 Discussion and Comparative Analysis

Robustness of Chamfer Similarity. As shown in Table 6, chamfer distance consistently yields the highest retrieval accuracy across all benchmarks. Its effectiveness is particularly pronounced in the K-pop Lyric Translation (Kim et al., 2024) and Covers80 (Ellis, 2007) datasets. By finding the best counterpart for every line in both the query and candidate songs, Chamfer distance remains robust to common variations in cover songs, such as structural reordering or the omission of specific verses.

F Detail in Cover Song Identification

Table 11 presents qualitative example of cross-lingual line-level matching results. The model is trained using the singability-aware framework with an E5-large backbone, and cosine similarity is used to measure cross-lingual line correspondence. For evaluation, we select K-pop songs that have both original Korean lyrics and corresponding English-version lyrics. As shown in the example, the model successfully aligns semantically and rhythmically corresponding lines across languages, even when literal translations are absent, while maintaining consistent syllabic structure, which reflects the singability-aware training objective.

Use of AI Assistants. During the preparation of this manuscript and associated visual materials, we utilized multiple large language model–based AI assistants, including ChatGPT (OpenAI), Claude (Anthropic), and Gemini (Google), as auxiliary tools. These tools were used to support English translation, grammar checking, and stylistic refinement of the writing, as well as to assist in drafting and refining visualizations used in the project. All scientific decisions, experimental designs, analyses, and interpretations were conducted and verified by the authors.

Section #1		
Transcription (Whisper-large-v3): I was a ghost, I was alone 아, 더워진 아, 기속에 Given the throne, I didn't know How to believe I was the queen that I meant to be I lived two lives Tried to play both sides But I couldn't find my own Caught a bad child Cause I got too wild But now that's how I'm getting paid		
Line	**GT**	**Predicted**
1	I was a ghost, I was alone, hah	I was a ghost, I was alone
2	어두워진, hah, 앞길속에 (Ah)	아, 더워진 아, 기속에
3	Given the throne, I didn't know how to believe	Given the throne, I didn't know How to believe
4	I was the queen that I'm meant to be	I was the queen that I meant to be
5	I lived two lives, tried to play both sides	I lived two lives Tried to play both sides
6	But I couldn't find my own place	But I couldn't find my own
7	Called a problem child 'cause I got too wild	Caught a bad child Cause I got too wild
8	But now that's how I'm getting paid, 끝없이 on stage	But now that's how I'm getting paid
Section #12		
Transcription (Whisper-large-v3): Waited so long to break these walls down To wake up and feel like me Put these patterns all in the past now And finally live like the girl they all see		
Line	**GT**	**Predicted**
1	Waited so long to break these walls down	Waited so long to break these walls down
2	To wake up and feel like me	To wake up and feel like me
3	Put these patterns all in the past now	Put these patterns all in the past now
4	And finally live like the girl they all see	And finally live like the girl they all see

Table 7: Genre: Animation Movie. Example of line-level segmentation results for the K-pop song "Golden" by HUNTR/X (from K-pop Demon Hunters).

Section #2		
Transcription (Whisper-large-v3): Shoes on, gotta bring them on Cup of milk, let's rock and roll King Kong, kick the drum Rollin' on like a rolling stone Sing song when I'm walkin' home Jump up to the table, LeBron Ding dong, call me on my phone Nice tea and I'll get my ping-pong		
Line	**GT**	**Predicted**
1	Shoes on, get up in the morn'	Shoes on, gotta bring them on
2	Cup of milk, let's rock and roll	Cup of milk, let's rock and roll
3	King Kong, kick the drum, rolling on like a Rolling Stone	Cup of milk, let's rock and roll Rollin' on like a rolling stone
4	Sing song when I'm walking home	Sing song when I'm walkin' home
5	Jump up to the top, LeBron	Jump up to the table, LeBron
6	Ding dong, call me on my phone	Ding dong, call me on my phone
7	Ice tea and a game of ping pong, huh	Nice tea and I'll get my ping-pong

Table 8: Genre: K-pop. Example of line-level segmentation results for 'Dynamite" by BTS.

Section #4		
Transcription (Whisper-large-v3): Only reason we are here is to celebrate In a place where anyone can be anything Hold on to this moment, don't let it fade away Baby, keep the music playing		
Line	**GT**	**Predicted**
1	Hey-oh, hey	
2	Only reason we are here is to celebrate	Only reason we are here is to celebrate
3	In a place where anyone can be anything	In a place where anyone can be anything
4	Hold on to this moment, don't let it fade away	Hold on to this moment, don't let it fade away
5	Baby, keep the music playing	Baby, keep the music playing
Section #5		
Transcription (Whisper-large-v3): Come on, get on up We're wild and we can't be tamed And we're turning the floor into a zoo Come on, keep it up It's fun if you're down to play And we're turning the floor into a zoo		
Line	**GT**	**Predicted**
1	Come on, get on up	Come on, get on up
2	We're wild and we can't be tamed	We're wild and we can't be tamed
3	And we're turnin' the floor into a zoo, ooh-ooh	And we're turning the floor into a zoo
4	Come on, keep it up	Come on, keep it up
5	It's fun if you're down to play	It's fun if you're down to play
6	And we're turnin' the floor into a zoo, ooh-ooh	And we're turning the floor into a zoo
Section #11		
Transcription (Whisper-large-v3): I'll take you higher, take you higher We can be ten, baby, I'll take you higher I'll take you higher, I'll take you higher We can be ten, baby, I'll take you higher		
Line	**GT**	**Predicted**
1	I'll take you higher, I'll take you higher	I'll take you higher, take you higher
2	We can't be tamed, baby, I'll take you higher	We can be ten, baby, I'll take you higher
3	I'll take you higher, I'll take you higher	I'll take you higher, I'll take you higher
4	And we can't be tamed, baby, I'll take you higher	We can be ten, baby, I'll take you higher
Section #12		
Transcription (Whisper-large-v3): Es una fiesta que sube como la espuma Yo partiré hasta la luna de ida y vuelta Es una fiesta que sube como la espuma Yo partiré hasta la luna de ida y vuelta		
Line	**GT**	**Predicted**
1	Es una fiesta que sube como la espuma	Es una fiesta que sube como la espuma
2	Yo por ti iré hasta la luna de ida y vuelta	Yo partiré hasta la luna de ida y vuelta
3	Es una fiesta que sube como la espuma	Es una fiesta que sube como la espuma
4	Yo por ti iré hasta la luna de ida y vuelta	Yo partiré hasta la luna de ida y vuelta

Table 9: Genre: Animation. Example of line-level segmentation results for the English lyrics of "Zoo" From Zootopia2 movie.

Section #6		
Transcription (Whisper-large-v3): It's whatever, it's whatever, it's whatever you like Turn this apatite into a club I'm talking drink, dance, smoke, freak, party all night Come on! Gumbay, gumbay, girl, what's up? Don't you want me like I want you, baby Don't you need me like I need you now Sleep tomorrow but tonight go crazy All you gotta do is just meet me at the		
Line	**GT**	**Predicted**
1	It's whatever it's whatever it's whatever you like	It's whatever, it's whatever, it's whatever you like
2	Turn this 아파트 into a club	Turn this apatite into a club
3	I'm talking drink, dance, smoke, freak, party all night	I'm talking drink, dance, smoke, freak, party all night Come on!
4	건배 건배 girl what's up	Gumbay, gumbay, girl, what's up?
5	Oh oh oh	
6	Don't you want me like I want you, baby	Don't you want me like I want you, baby
7	Don't you need me like I need you now	Don't you need me like I need you now
8	Sleep tomorrow but tonight go crazy	Sleep tomorrow but tonight go crazy
9	All you gotta do is just meet me at the	All you gotta do is just meet me at the
Section #12		
Transcription (Whisper-large-v3): Hey so now u know the game Are u ready? Cuz im coming to get u Get u get u Hold on		
Line	**GT**	**Predicted**
1	Hey so now you know the game	Hey so now u know the game
2	Are you ready?	Are u ready?
3	Cause I'm comin to get ya	Cuz im coming to get u
4	Get ya, get ya	Get u get u
5	Hold on, hold on	Hold on
6	I'm on my way	
7	Yeah yeah yeah yeah yeah	
8	I'm on my way	

Table 10: Genre: K-pop. Example of line-level segmentation results for "APT." by ROSÉ and Bruno Mars.

Line	Sim	Korean Lyrics (syll.)	English Lyrics (syll.)
1	1.0000	I know I want it (5)	I know I want it (5)
2	0.3701	입에 바른 소린 이제 그만할게 (12)	I don't wanna hide, pretending I don't want it (14)
3	1.0000	'Cause I deserve it (Deserve it) (8)	'Cause I deserve it (Deserve it) (8)
4	0.6137	혹시 잠깐 내가 미워지더라도 걱정 안 할게 (17)	I'm not worried What you'll think of me when you see what's inside me (19)
5	1.0000	'Cause I know you (I know you) (7)	'Cause I know you (I know you) (7)
6	0.7632	내 눈을 자꾸 피해봐 (Hey) (9)	Don't look away, look in my eyes (Hey) (10)
7	0.8350	네 맘을 자꾸 숨겨봐 (Hey) (9)	Show me your heart, do not disguise (Hey) (9)
8	0.7189	나에게서 도망쳐봐, no, no (10)	You can run, but you cannot hide, no, no (10)
9	0.5562	감았던 눈을 떴을 때 (Hey) (9)	Even if you tell me goodbye (Hey) (8)
10	0.7508	문득 내가 떠오를 때 (Hey) (9)	Whenever you open your eyes (Hey) (9)
11	0.5065	You are gonna be mine again, yeah (9)	You'll think of me, you're hypnotized (Yeah, yeah) (13)
12	0.9971	You're gonna say more, more, more, more, more, and more (12)	You're gonna say more, more, more, more, more and more (12)
13	0.8653	멈추지 못해 more, more, more, and more (10)	Can't help but say more, more, more, and more (10)
14	0.4095	그러니 한 번 더 (6)	'Cause you can't get enough (7)
15	0.9967	I wanna have more, more, more, more, more, and more (11)	I wanna have more, more, more, more, more and more (11)
16	0.7169	(멈추기 싫어 more, more, more, and more) (10)	(And you will say more, more, more, and more) (9)
17	0.8248	(그러니 한 번 더) More (7)	('Cause it's never enough) More (8)
18	0.3429	멈추지를 못해 (6)	You'll be begging for more (7)
19	1.0000	(More and more) (3)	(More and more) (3)
20	0.5805	그러니 한 번 더 더 (7)	'Cause you can't get enough, -nough (8)
21	0.2742	멈추기가 싫어 (6)	You'll be craving for more (7)
22	1.0000	(More and more) (3)	(More and more) (3)
23	0.6427	그러니 한 번 더 (6)	'Cause it's never enough (7)
24	1.0000	Do you feel me? (4)	Do you feel me? (4)
25	0.6163	니가 날 위한 사람이라고 믿니? (12)	Or do you still not understand how much you love me? (13)
26	0.5656	Only for me (Only for me) (8)	Boy, you need me, you want me (7)
27	0.4857	간지러운 말은 굳이 하지 않아도 넌 (14)	Stop pretending I can see it in your eyes, you're already mine (17)
28	0.8334	'Cause you know me, you know me (7)	'Cause I know you, I know you (7)
29	0.6235	네 귀를 자꾸 막아도 (8)	Even if you cover your ears (Ah) (9)
30	0.5306	나를 멀리 밀어내도 (8)	Push me away for years and years (Ah) (9)
31	0.7571	나에게서 멀어져도 (저 멀리), no, no (13)	In your heart, I won't disappear (Be right there), no, no (14)
32	0.6155	내가 다시 널 부르면 (8)	If I ever call out your name (Ah) (9)
33	0.1410	나의 목소릴 들으면 (8)	Your heart will feel just like the same (Ah) (9)
34	1.0000	You are gonna be mine again (8)	You are gonna be mine again (8)
35	0.6757	Yeah, 한 번 더 (4)	And once again! (4)
36	0.9971	You're gonna say more, more, more, more, more, and more (12)	You're gonna say more, more, more, more, more and more (12)
37	0.8653	멈추지 못해 more, more, more, and more (10)	Can't help but say more, more, more, and more (10)
38	0.4095	그러니 한 번 더 (6)	'Cause you can't get enough (7)
39	0.9967	I wanna have more, more, more, more, more, and more (11)	I wanna have more, more, more, more, more and more (11)
40	0.7169	(멈추기 싫어 more, more, more, and more) (10)	(And you will say more, more, more, and more) (9)
41	0.8248	(그러니 한 번 더) More (7)	('Cause it's never enough) More (8)
42	0.3429	멈추지를 못해 (6)	You'll be begging for more (7)
43	1.0000	(More and more) (3)	(More and more) (3)
44	0.5805	그러니 한 번 더 더 (7)	'Cause you can't get enough, -nough (8)
45	0.2742	멈추기가 싫어 (6)	You'll be craving for more (7)
46	1.0000	(More and more) (3)	(More and more) (3)
47	0.6427	그러니 한 번 더 (6)	'Cause it's never enough (7)
48	0.7586	난 원래 욕심쟁이 몰랐다면 미안 (13)	I'm a greedy girl, sorry if you didn't know (13)
49	0.8648	사과는 미리 할게 'cause I want you more, more (13)	Take my apology because I want you more and more (14)
50	0.6892	의견은 필요 없어 훔칠 거야 네 맘 (13)	Gonna steal your heart, own it and won't let it go (
,13)			
51	0.8339	내게 홀리게 될걸 you can't say no, no (13)	By then you'll understand you can't say no, no (13)
52	0.7697	난 도둑고양이 오늘만은 널 (11)	I'll be sneaking in your heart like a stray cat (12)
53	0.2339	꼭 잡으러 왔으니까 딱 기다려 너 (13)	Conquer it and put my mark on it like a tat (12)
54	0.4545	멀리 가지 말고 다시 내게 come, come (12)	My love, so sweet the plan will never fall flat (11)
55	0.7981	Yeah 한 번 더 (More) (5)	And once again! (More) (5)
56	0.7065	멈추지를 못해 (You're gonna say) (11)	You'll be begging for more (You're gonna say) (12)
57	1.0000	(More and more) (3)	(More and more) (3)
58	0.6876	그러니 한 번 더 더 (그러니 한 번 더) (13)	'Cause you can't get enough, -nough ('Cause it's never enough) (15)
59	0.7186	멈추기가 싫어 (More and more) (9)	You'll be craving for more (More and more) (10)
60	1.0000	(More and more) (3)	(More and more) (3)
61	0.8688	그러니 한 번 더 (More) (7)	'Cause it's never enough (More) (8)

Table 11: Line-level cross-lingual matching examples for *all lyric lines* of the song "More&More" by TWICE, with cosine similarity (Sim) and syllable counts (syll.). Genre: K-pop.