Towards Singable Lyrics Translation Using Large Language Models

Hanze Liu, Yusuke Sakai, Taro Watanabe
Nara Institute of Science and Technology (NAIST), Japan
liu.hanze.li9@naist.ac.jp
sakai.yusuke.sr9@is.naist.jp
taro@is.naist.jp

Abstract

Lyrics translation must account for rhythm, rhyme, and singability in the translated lyrics. In this study, we focus on singability and investigate effective prompting methods for translating singable lyrics, including verification-guided and multi-round prompting, applied to large language models. First, we curate a multilingual lyrics translation dataset covering a total of six language directions across Chinese, Japanese, and English. Next, we evaluate seven prompting strategies, with instruction complexity increasing incrementally. The results show that multi-prompt strategies improve singability-related aspects, such as rhythmic alignment and phonological naturalness, compared to naive translation. Furthermore, human evaluations using songs created from translated lyrics suggest that moderately complex prompting strategies improve singable naturalness, while more complex strategies contribute to greater stability in perceived quality.

1 Introduction

Lyrics translation is a form of constrained translation that extends beyond conventional machine translation tasks. It requires balancing linguistic fidelity with musical form, capturing not only literal meaning but also rhythm and rhyme to achieve overall singability. Accordingly, effective lyrics translation must explicitly account for rhythm, rhyme, and singability (Low, 2003, 2005).

Such lyrics translation has become increasingly important with the growth of subcultural platforms such as karaoke services and video streaming websites, as well as the globalization of the music market. In these contexts, immediacy directly affects a song's popularity and playback counts, making automatic lyrics translation a crucial component of music distribution. Moreover, translations that are easy to hum and easy to understand in the target language are often preferred. Consequently, effective lyrics translation often prioritizes paraphrasing and phonologically informed language fitting over strict literal translation, as illustrated in Figure 1.

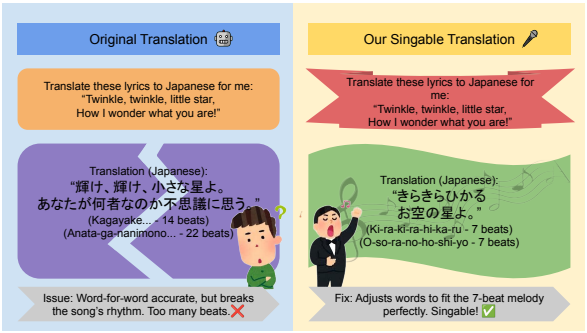

Figure 1: Overview of singable lyrics translation. Lyrics translation must consider not only semantic accuracy but also melodic fit to ensure singability. Furthermore, these aspects are difficult to evaluate using automatic metrics alone, making human perception evaluation based on songs created from the translated lyrics essential.

Some existing work on lyrics translation has primarily relied on fine-tuning open-source language models for this domain (Ou et al., 2023). In contrast, research on zero-shot prompting strategies for lyrics translation has been relatively limited, with most prior efforts focusing on model development or dataset construction rather than systematic exploration of prompting strategies (Cho et al., 2025). Related techniques have been applied to poetry translation, which constitutes a similar form of constrained translation (Song et al., 2023). While poetry translation likewise demands a high level of textual aesthetics, the requirements for musicality and singability, which are less directly tied to wording yet play a more critical role in lyrics quality, are comparatively lower than in lyrics translation.

In this study, we explore effective prompting strategies for lyrics translation in a zero-shot setting using large language models (LLMs). We propose a set of prompts for a multilingual lyrics translation task and comprehensively evaluate their effects using both linguistic metrics and music-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 544–554

related metrics. Specifically, we design a series of prompts with increasing levels of complexity and specificity and examine their impact on semantic similarity, syllable or mora-based rhythm, rhyme structure alignment, and phonotactic difficulty. In addition, through human evaluation, we analyze perceptual differences that automatic evaluation cannot fully capture. Finally, based on these human judgments, we investigate which automatic metrics show the strongest correlation with human perception of singability, thereby providing insights into more reliable evaluation of singable lyrics translation. Our contributions and findings are as follows:

- We conduct a comprehensive investigation of the zero-shot capabilities of large language models for lyrics translation by designing seven prompting strategies with progressively increasing complexity, including multi-prompt strategies and verification-guided prompting.
- We manually construct a multilingual lyrics translation dataset aligned across Japanese, Chinese, and English, annotated with syllable or mora counts for evaluation. Using this, we conduct comprehensive evaluations and demonstrate that appropriate prompting strategies help maintain consistency across singability-related aspects.
- We manually create actual songs using translated lyrics and conduct human evaluation. The results indicate that prompting strategies with moderate complexity achieve sufficient perceptual improvements, while more complex strategies, such as multi-prompt prompting, tend to yield more stable results when considering score variance.
- We conduct a meta-evaluation using these human-evaluation results to assess their correlation with automatic metrics and find that CCVO shows the strongest alignment with human perception.

2 Background and Related Work

Lyrics Translation. Kim et al. (2023) propose an evaluation framework for singable lyrics translation that incorporates syllable counts, phoneme repetition, musical structure, and semantic similarity. Štěpánková and Rosa (2025) provides a computational interpretation of the Pentathlon Principle, introducing measurable, music-aware metrics, including rhyme- and phonology-focused measures. Complementarily, Ou et al. (2023) treats lyrics translation as a form of constrained machine translation, using prompt-based controls over properties such as length and rhyme. Ye et al. (2024) introduces a method for translating musical lyrics, emphasizing the balance between semantic fidelity and singability. Lyrics translation is often performed with respect to the structural organization of songs, such as lyrical sections within a part or sentence-level units corresponding to individual musical bars. These studies suggest that the Pentathlon Principle has become widely supported for lyrics translation. Therefore, evaluation is expected to consider multiple aspects, including singability, meaning preservation, naturalness, rhythm, and rhyme. However, despite this shared perspective, there remain few well-established evaluation metrics, making human evaluation particularly important for assessing lyrics translation quality.

Figure 2: Overview of our prompting strategies. We are considering three aspects, a total of seven prompts.

Prompting. For creative domains, Wang et al. (2024) systematically analyzes performance in English to Chinese poetry translation under different prompting settings. Similarly, Pramodya et al. (2025) conduct a comprehensive study of the effects of prompting strategies in movie translation. For lyric translation, Cho et al. (2025) partially adopts zero-shot prompting with LLMs. However, their work primarily focuses on dataset construction and presents only preliminary model comparisons, without investigating prompting strategies. Prompting has been shown to enhance the effectiveness of large language models in zero-shot settings (Kojima et al., 2022), while iterative self-feedback improves generation quality (Madaan et al., 2023), and zero-shot verification approaches further strengthen model reliability (Miao et al., 2024). Accordingly, these techniques suggest strong potential for application to lyrics translation.

3 Methodology

To explore effective prompting strategies for lyrics translation, we conduct a systematic investigation from three perspectives, as illustrated in Figure 2: context-aware prompting, verification steps, and multi-round prompting. This prompt design allows us to analyze the effects of prompt complexity and

structural control on lyrics translation. Note that the target lyrics are provided as a user-prompt following the instruction prompt. Therefore, we show only the instruction-prompt component.

Experiment 1: Simple prompt. Exp. 1 serves as our zero-shot baseline, in which the model is instructed to directly translate the source lyrics into the target-language lyrics. This prompt contains only the core task instruction and no additional constraints related to singability.

Experiment 1: Simple prompt

Translate this passage of lyrics into {target_language} lyrics.

Experiment 2: Simple prompt with minimal constraint. Exp. 2 introduces a requirement for line-by-line alignment between the source lyrics and the translation. This setting enables us to examine whether adding a basic structural instruction improves performance and to what extent models can conform to explicit instructions.

Experiment 2: Simple prompt with minimal constraint

Translate this passage of lyrics into {target_language}, **line by line**.

Experiment 3: Constraint-aware prompt. As opposed to Exp. 2, Exp. 3 introduces multiple explicit structural constraints to test whether models can simultaneously conform to these objectives. Specifically, Exp. 3 requires the model to verify rhythm (mora count or syllable count), rhyme patterns, and semantic consistency for each lyric.

Experiment 3: Constraint-aware prompt

Translate the passage line by line into {target_language} **with: 1) same mora count per line; 2) same rhyme pattern; 3) same theme**.

Experiment 4: Enhanced constraint-aware prompt. Exp. 4 extends the constraints introduced in Exp. 3 by incorporating an explicit verification mechanism alongside the requirements. Rather than merely stating the constraints, we specify how the model should verify compliance, including mora or syllable counting for rhythm, phoneme analysis for rhyme, and thematic preservation checks for meaning, drawing inspiration from step-by-step reasoning approaches (Miao et al., 2024). Through this design, we test whether verification-guided instructions improve performance in translating singable lyrics.

Exp. 4: Enhanced constraint-aware prompt

Translate the passage line by line into {target_language} with 1) keep the same mora count per line and **verify by counting mora in hiragana/syllables**; 2) keep the same rhyme pattern **and verify by the final phoneme per line**; 3) keep the same theme.

Experiment 5: Enhanced constraint-aware prompt with general refinement (two-turn). Exp. 5 introduces a two-turn interaction by extending Exp.4's enhanced constraint-aware approach with a refinement process. After generating an initial translation using Exp.4's constraints, the model will received a second-turn instruction to improve translation quality without re-stating specific constraints. This design is inspired by Madaan et al. (2023) and investigates if the model can implicitly maintain previously established constraints while optimizing for better results.

Experiment 5: Enhanced constraint-aware prompt with general refinement (two-turn)

Round 1: Translate the passage line by line into {target_language} with 1) keep the same mora count per line and verify by counting mora in hiragana/syllables; 2) keep the same rhyme pattern and verify by the final phoneme per line; 3) keep the same theme.

Round 2: Without changing the original requirements, refine the translation for better overall translation quality.

Experiment 6: Enhanced constraint-aware prompt with rhythm refinement (two-turn). While Exp. 5 employs general refinement in the

second round, Exp. 6 implements targeted refinements focusing on mora or syllable counts. In the first round, Exp. 6 follows the same instructions as Exp. 5. In the second round, Exp. 6 focuses explicitly on verifying mora or syllable counts. This design tests whether constraint-specific refinements can improve overall translation quality, given the importance of rhythm for singability, especially in fitting translated lyrics to the original melody.

Experiment 6: Enhanced constraint-aware prompt with rhythm refinement (two-turn)

Round 1: Translate the passage line by line into {target_language} with 1) keep the same mora count per line and verify by counting mora in hiragana/syllables; 2) keep the same rhyme pattern and verify by the final phoneme per line; 3) keep the same theme.

Round 2: Without changing meaning, adjust so each line keeps the same mora/syllable count; verify via syllable counting.

Experiment 7: Enhanced constraint-aware prompt with rhythm refinement (multi-turn). Finally, instead of focusing on individual sections only, Exp. 7 shifts the scope to song-level translation by keeping the translation of all sections within a unified conversation, while still maintaining a section-by-section translation procedure. In Exp. 7, the model translates all sections sequentially within a single dialogue, applying the constraint-aware prompt from Exp. 4 at each turn. This design tests whether accumulated conversational context improves consistency and coherence in singable lyrics translation, addressing potential fragmentation that may arise when sections are translated in isolation.

Experiment 7: Enhanced constraint-aware prompt with rhythm refinement (multi-turn)

All rounds: Translate the passage line by line into {target_language} with 1) keep the same mora count per line and verify by counting mora in hiragana/syllables; 2) keep the same rhyme pattern and verify by the final phoneme per line; 3) keep the same theme.

4 Experimental Setup

4.1 Evaluation Dataset and Preprocessing

To enable lyrics translation in a multilingual setting, we curate a dataset in which lyrics are mutually annotated across Chinese, Japanese, and English. We construct a multilingual lyrics dataset segmented by lyrical sections. The dataset consists of 96 translated songs across the three languages, comprising a total of 624 sections. All data are manually transcribed from social media platforms such as YouTube[1] and Bilibili[2]. Each song is associated with at least one valid audio source, ensuring that both the original and translated lyrics are singable by human singers. We primarily transcribe the lyrics from the song audio, using expressions from comment sections or on-screen text in the videos as additional references when necessary to complete the transcription. To facilitate rhythm-aware evaluation, we manually counted language-specific rhythmic units: pinyin-based syllable counts for Chinese, mora counts derived from hiragana for Japanese, and computational syllabification for English. Using this, we evaluate all six possible translation directions among the three languages.

4.2 Translation Settings

We compare seven prompting strategies introduced in Section 3. To capture the effect of prompting strategies, we use a single model, enabling a direct comparison of pure generation results attributable solely to differences in prompting strategies. All translations are generated using the `GPT-4o` (gpt-4o-2024-11-20) (OpenAI et al., 2024) under a zero-shot prompting setting. To ensure clean outputs as the model may response to prompts designed to guide its internal reasoning, we apply a post-processing step using a second API call to remove any extraneous content, with a manual checking process to verify the processed lyrics, ensuring that only the translated lyrics remain for evaluation. To ensure consistency across experiments, decoding parameters are fixed ($temperature = 0$, $top\_p = 0$), resulting in deterministic generation. remains deterministic. Following Kim et al. (2023), model inputs are provided at the section level instead of on a line-by-line basis, since line-level translation fails to preserve important contextual information within a section.

[1] https://www.youtube.com/
[2] https://www.bilibili.com/

4.3 Evaluation Metrics

To explore prompting strategies optimized for singable lyrics translation, we consider two automatic evaluation aspects: translation quality, and lyrics-aware naturalness and fitness for singability.

4.3.1 Translation Quality

For translation quality, we employ three neural evaluation metrics. Since lyrics translation often prefers interpretative translation rather than strict literal translation, surface-level metrics such as BLEU (Papineni et al., 2002) or chrF (Popović, 2015) are less suitable. Instead, we evaluate translation quality using semantics-aware neural metrics that better capture meaning preservation.

Sentence-BERT (Reimers and Gurevych, 2019) It measures semantic similarity at the sentence or line level, making it suitable for capturing paraphrastic alignment. By comparing the generated lyrics with the reference translations, the metric captures alignment with human translations. For English-target translation directions, we use `all-MiniLM-L12-v2`[3] as the backbone model. For non-English directions, we use `distiluse-base-multilingual-cased-v2`[4].

COMET (Rei et al., 2020) A neural evaluation metric that estimates translation quality by jointly modeling adequacy and fluency, using source sentences, reference translations, and generated lyrics to assess semantic alignment and linguistic naturalness. We used `Unbabel/wmt22-comet-da`[5].

COMET-QE (Rei et al., 2022) A reference-free quality estimation metric that predicts translation quality based solely on the source sentence and the generated output, allowing meaning preservation to be assessed without reference translations. We use `Unbabel/wmt22-cometkiwi-da`[6].

4.3.2 Lyrics-Specific Evaluation

For lyrics-specific evaluation, we followed the Pentathlon Principle (Low, 2003, 2005; Štěpánková and Rosa, 2025). According to the Pentathlon Principle, effective lyrics translation must account for several aspects, including *singability*, e.g., matching the melody, *sense*, e.g., semantic accuracy, *naturalness*, e.g., idiomaticity in the target language, *rhyme*, e.g., preserving sound patterns, and *rhythm*, e.g., matching syllable counts and stress. Based on this principle, we focus on lyrics-specific singability aspects that are not fully captured by standard translation quality metrics. We evaluate these aspects using three lyrics-aware metrics: rhythm, rhyme, and CCVO (Consonant Cluster and Vowel Openness Distance).

Rhythm (syllable distance) (Kim et al., 2023). We define the rhythm distance between a source sequence $X = (x_1, \ldots, x_n)$ and a translation $\tilde{X} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ as:

$$\mathrm{Dis}_{\mathrm{syl}}(X, \tilde{X}) = \frac{1}{2n} \sum_{i=1}^{n} \left(\frac{|\mathrm{syl}(x_i) - \mathrm{syl}(\tilde{x}_i)|}{\mathrm{syl}(x_i)} + \frac{|\mathrm{syl}(x_i) - \mathrm{syl}(\tilde{x}_i)|}{\mathrm{syl}(\tilde{x}_i)} \right),$$

where $\mathrm{syl}(\cdot)$ denotes the number of syllables (or mora for Japanese). Lower values indicate better rhythmic alignment. It encourages rhythmic consistency between the source and translated lyrics, and translations with lower distance are considered more singable, as they can be more easily aligned with the original melody and musical phrasing.

Rhyme similarity (Štěpánková and Rosa, 2025) We compute rhyme structure similarity as the Jaccard index between sets of rhyming edges:

$$RS_{JI}(R, \tilde{R}) = \frac{|Edges(R) \cap Edges(\tilde{R})|}{|Edges(R) \cup Edges(\tilde{R})|},$$

where $Edges(R)$ denotes the set of line pairs in the reference R that share the same rhyme class, and $\mathrm{Edges}(\tilde{R})$ is defined analogously for the translated lyrics. Higher values indicate greater similarity in rhyme structure between the reference and translated lyrics, reflecting better preservation of rhyming patterns.

CCVO (Consonant Cluster and Vowel Openness Distance) (Štěpánková and Rosa, 2025) To quantify phonological singability, we define the CCVO distance as the average normalized Levenshtein distance between CCVO encodings of corresponding lines:

$$CCVO_{\mathrm{Dist}}(X, \tilde{X}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{Lev}_{\mathrm{Dist}}(CCVO(x_i), CCVO(\tilde{x}_i))}{\mathrm{len}(CCVO(x_i))},$$

where $CCVO(\cdot)$ encodes each line into a sequence representing consonant cluster complexity and vowel openness classes. Text from Chinese, English, and Japanese is mapped to IPA form to

[3] https://hf.co/sentence-transformers/all-MiniLM-L12-v2

[4] https://hf.co/sentence-transformers/distiluse-base-multilingual-cased-v2

[5] https://hf.co/Unbabel/wmt22-comet-da

[6] https://hf.co/Unbabel/wmt22-cometkiwi-da

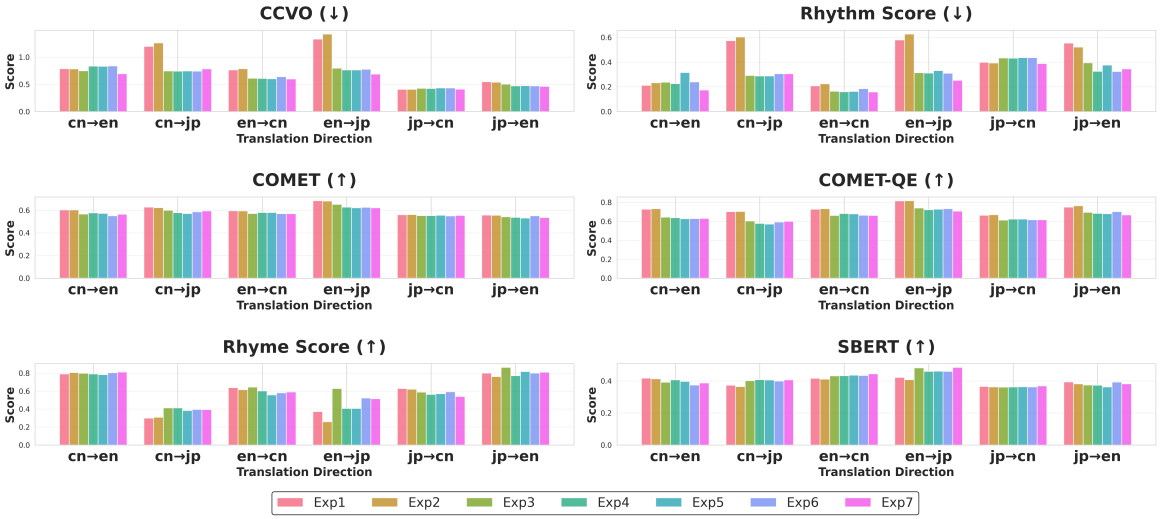

Figure 3: Evaluation results with six evaluation metrics, six language directions, and seven prompting settings, where Exp. 1 uses a simple prompt, Exp. 2 incorporates minimal constraints, Exp. 3 adopts a constraint-aware prompt, Exp. 4 further enhances it, Exp. 5 introduces two-turn refinement, Exp. 6 applies two-turn rhythm refinement, and Exp. 7 extends it to multi-turn rhythm refinement.

extract phonemes. We then label consonant clusters, which are determined by whether there are 3 or more consecutive consonants across syllable boundaries, and label vowel openness into 3 categories: open, mid, and closed. Using a string composed of consonant cluster and vowel openness labels, we compute the normalized Levenshtein distance between CCVO strings, normalized by source sequence length. Lower scores indicate closer phonological profiles between the source and translated lyrics, reflecting higher phonological compatibility for singing. We primarily adopt this metric because it is more directly related to singability (Štěpánková and Rosa, 2025).

5 Experimental Results

Figure 3 shows the overall evaluation results across all six translation directions. We discuss the key observations and aspects based on these results.

Finding 1: Complex prompting strategies improve singability in translations particularly into Japanese. CCVO and rhythm scores remain relatively stable across different levels of prompt complexity for the CN→EN, EN→CN, JP→CN, and JP→EN directions. In contrast, performance degrades with simpler prompting strategies for the CN→JP and EN→JP directions. This indicates that naive prompting strategies struggle to properly align Japanese mora with syllable-based representations, whereas more advanced prompting strategies, such as iterative prompting, substantially improve phonological alignment. Notably, a moderate level of prompt complexity, e.g., Exp. 3, is sufficient to achieve most of the gains in singability-related metrics. While higher complexity settings, e.g., Exp. 7, lead to further improvements in CCVO, these gains are mainly observed in non-Japanese target directions such as CN→EN.

Finding 2: Translation quality appears stable under automatic evaluation, and complex prompting may show higher correlation with human translations. Across COMET, COMET-QE, and Sentence-BERT (SBERT), translation quality scores remain largely stable in most settings, indicating that automatic evaluation metrics are relatively insensitive to differences in prompting strategies. One exception is COMET-QE, which relies solely on the source lyrics and the generated translation without reference translations. As a result, simpler and more literal prompting strategies, e.g., Exp. 1, tend to receive slightly higher scores under COMET-QE. In contrast, different trends emerge when comparing SBERT and COMET-QE, particularly for the EN→JP direction. SBERT shows higher similarity to human reference translations under more complex prompting strategies, suggesting that these prompts enable translations that deviate from strictly literal translation and better capture culturally informed paraphrasing. This observation indicates that, although automatic metrics may appear stable overall, more complex prompting tends to produce outputs that align more

Exp.1	Exp.7	Reference	Original
你愿加入我们的征途吗 (10) Will you join in our journey	你会加入我们的战斗吗 (10) Will you join in our battle	你会加入正义军吗 (8) Will you join in our righteous army	Will you join in our crusade (7)
谁会坚强地与我并肩站立 (11) Who will firmly stand side-by-side with me	谁会坚强地与我同行 (9) Who will firmly walk alongside with me?	与我并肩去作战 (7) Side-by-side with me to battle	Who will be strong and stand with me (8)
在那路障之外 (6) Beyond the roadblock	在那街垒之外 (6) Beyond the barricades	用血肉筑起街垒 (7) Construct the barricade with flesh and blood	Beyond the barricade (6)
是否有你渴望的世界 (9) Is there a world that you long for?	是否有你渴望的天地 (9) Is there a realm that you long for?	为那理想共患难 (7) Sharing trials and tribulations for that ideal.	Is there a world you long to see (8)

Table 1: The example from "*Do you hear the people sing?*", English-to-Chinese translation. Numbers in parentheses denote syllable counts. Exp.1: Experiment 1 output (rhythm 0.2438, COMET 0.6302, COMET-QE 0.7302). Exp.7: Experiment 7 output (rhythm 0.1461, COMET 0.6788, COMET-QE 0.7798). Reference: A singable version verified through human vocal performance Original: Original English version.

closely with human translation preferences.

Finding 3: Prompting enhances singability and rhyme awareness while maintaining translation quality. Focusing on rhyme scores, we observe greater variability across translation directions, particularly for EN→JP, JP→EN, and CN→JP. Nevertheless, rhyme similarity consistently improves with prompting strategies of moderate complexity or higher, e.g. Exp. 3, compared to simpler prompting such as Exp. 1. In contrast to CCVO and rhythm scores, rhyme scores exhibit a somewhat different trend, and improvements in rhyme do not necessarily lead to substantial gains in the other two phonological metrics. When considering translation quality metrics such as SBERT, moderately complex prompting, e.g., Exp. 3, appears effective in several cases. Nevertheless, increasing prompt complexity tends to result in more consistent improvements across multiple evaluation metrics, while largely preserving translation quality.

6 Qualitative Analysis

As discussed in Section 3, lyrics translation often involves subtle variations that cannot be fully captured by automatic evaluation metrics. Nevertheless, certain tendencies can still be observed and are further examined through qualitative analysis. Table 1 illustrates this point with an example from the song "*Do You Hear the People Sing?*", translated from English into Chinese, showing that prompting strategies providing clear instructions can improve rhythmic alignment. In Exp. 1, the model receives only basic prompting instructions without explicit structural constraints. In contrast, Exp. 7 achieves improved rhythmic alignment and semantic similarity, while the original human reference represents an upper bound in terms of both semantic fidelity and rhythmic conformity. This comparison illustrates how explicit structural constraints in prompting can enhance musical form in lyrics translation.

Regarding translation quality, qualitative inspection reveals that the overall meaning remains largely consistent between Exp. 1 and Exp. 7. However, Exp. 7 tends to select terminology that is closer to the reference translation, particularly in finer-grained lexical choices. At the same time, the human reference sometimes departs from the source content; in the final line of the example, the reference preserves the original meaning less faithfully than the model outputs. These observations highlight the inherent variability and creativity involved in lyrics translation. Nevertheless, across cases, prompting strategies enable conservative and goal-oriented translation shifts without substantially altering the model's underlying creativity. This suggests that prompting can effectively guide lyrics translation in a controllable manner, balancing faithfulness and creative adaptation.

7 Human Evaluation

7.1 Settings

To further evaluate and validate translation quality from the perspective of human perception, we conduct a human evaluation with three native speakers per each target language direction. The evaluators

Prompt	MOS	Naturalness	Melody Fit	Emotion
Original	3.64 ± 1.07	3.66 ± 1.19	3.53 ± 1.13	3.78 ± 1.09
Exp. 1	3.11 ± 1.02	$\mathbf{3.36 \pm 1.24}$	3.06 ± 1.10	3.41 ± 0.86
Exp. 2	2.90 ± 1.13	3.08 ± 1.27	2.88 ± 1.15	3.17 ± 1.07
Exp. 3	3.25 ± 0.84	3.25 ± 0.98	3.40 ± 1.00	3.47 ± 0.84
Exp. 4	$\mathbf{3.30 \pm 0.88}$	3.35 ± 1.15	$\mathbf{3.48 \pm 1.00}$	3.44 ± 0.85
Exp. 5	3.04 ± 0.76	3.03 ± 1.07	3.28 ± 1.02	3.32 ± 0.81
Exp. 6	3.18 ± 0.81	3.15 ± 1.11	3.32 ± 0.93	3.39 ± 0.81
Exp. 7	3.22 ± 0.80	3.25 ± 1.04	3.25 ± 0.90	$\mathbf{3.50 \pm 0.78}$

Table 2: Aggregated Human Evaluation Results

rate the translated lyrics across multiple quality dimensions. For each target language, we randomly select 88 sections and manually create songs by combining the translated lyrics with the original melodies. To ensure consistency across samples, we adopt a uniform adjustment strategy: when the translated lyrics contain fewer syllables, we prolong musical notes, and when they contain more syllables, we map multiple syllables to a single note. This procedure ensures that all samples are produced under a consistent and controlled setting. The Evaluators assess translated lyrics on a 5-point scale (1=very poor, 5=excellent) across four dimensions: **1. Naturalness**: Fluency and idiomaticity of the lyrics; **2. Lyric-Melody Fit**: How well the translation fits the original melody when sung; **3. Emotional Impact**: Preservation of emotional content and artistic intent; and **4. MOS (Mean Opinion Score)**: Overall translation quality. We collected evaluations across 42 experiment combinations (7 prompts $\times$ 6 translation directions). Each rating represents the aggregated judgment of multiple human evaluators on sampled translations from the corresponding experiment and direction.

7.2 Results

Table 2 presents the aggregated results of the human evaluation. Prompt 4 achieved the highest overall quality (MOS: 3.30 ± 0.88), demonstrating the most consistent performance across all translation directions. In contrast, Prompt 2 received the lowest ratings (MOS: 2.90 ± 1.13), indicating that incomplete or underspecified instructions can degrade translation quality, in some cases resulting in worse performance than providing no instructions at all. Using Japanese as the source language led to significantly better translations (average MOS: 3.70 for JP$\rightarrow$CN and 3.40 for JP$\rightarrow$EN) compared to other language pairs. Conversely, translations into Chinese proved the most challenging, with CN-target experiments exhibiting the largest performance variance across prompting strategies. In addition, when focusing on score variance, we observe that iterative refinement strategies such as Exp. 7 improve the consistency of human judgments such as Emotion and Melody Fit. This suggests that repeated refinement not only improves average quality but also stabilizes perceived singability-related attributes.

7.3 Correlation with Automatic Metrics

We compute Pearson correlations between human ratings and automatic metrics across all 48 evaluation points to assess the validity of these metrics for lyrics translation. CCVO exhibits a very strong positive correlation with both MOS ($r = 0.838$, $p < 0.001$) and Naturalness ($r = 0.705$, $p < 0.001$). Translations with better phonological alignment consistently receive higher human ratings, supporting CCVO as a reliable automatic indicator of lyrics translation quality. In contrast, SBERT ($r = -0.437$, $p < 0.01$), COMET ($r = -0.755$, $p < 0.001$), and COMET-QE ($r = -0.464$, $p < 0.01$) show strong negative correlations with human judgments. In particular, higher COMET scores are associated with lower human ratings, suggesting that these semantics-oriented metrics are not well suited to evaluating the quality of singable lyrics translation. Rhyme similarity shows a moderate positive correlation with human ratings ($r = 0.340$ for MOS, $p < 0.05$; $r = 0.393$ for Naturalness, $p < 0.01$), indicating that preservation of rhyme patterns contributes to perceived translation quality. In contrast, rhythm exhibits no significant correlation with human ratings ($r = -0.161$, $p > 0.05$), suggesting that syllable-level alignment alone does not guarantee perceived quality. Overall, these results indicate that CCVO is the most informative automatic predictor of human-perceived quality among the metrics examined.

8 Conclusion

In this study, to investigate the zero-shot lyrics translation ability of large language models and identify effective prompting strategies, we design a total of seven prompting strategies with varying levels of complexity, incorporating techniques such as verification-guided prompting and multi-round prompting. We conducted a comprehensive evaluation of these strategies from multiple perspectives related to translation quality and singability.

We also curated a multilingual lyrics translation

dataset consisting of 96 translated songs (624 sections) across Chinese, Japanese, and English, with aligned translations across the three languages. Experimental results showed that complex prompting strategies improve singability, particularly for translations into Japanese. While translation quality remains largely stable under automatic evaluation, more complex prompting strategies tend to show higher alignment with human translations. Moreover, prompting enhances singability and rhyme awareness while preserving translation quality.

Human evaluation further reveals that prompting strategies with moderate complexity, e.g., Exp. 4, achieve the best perceived quality on average, whereas multi-round prompting can improve the stability and consistency of human judgments. Finally, by treating human evaluation results as a form of meta-evaluation and measuring their correlation with automatic metrics, we find that CCVO exhibits the strongest alignment with human perception, suggesting its potential as a proxy metric for human evaluation in singable lyrics translation.

We believe that these insights provide useful directions for future work, such as exploring melody-aware prompting, and developing lyrics-specific evaluation metrics that better connect semantic similarity with creative translation quality.

Limitations

LLMs. Since this study focuses on prompting strategies, all experiments are conducted using a single large language model. While evaluating multiple LLMs could increase the comprehensiveness of the analysis, our primary contribution lies not in cross-model performance comparison, but in examining how variations in prompting strategies affect singability in lyrics translation. Accordingly, we do not explore LLM variation in this work. Instead, by restricting our analysis to a single model, we can provide a more in-depth and controlled investigation of prompting effects. Moreover, several prior studies similarly adopt a single LLM setting to enable deeper analysis of prompting behavior (Wang et al., 2024; Hu et al., 2024; Yang et al., 2024). For this reason, multi-model evaluation is treated as a nice-to-have future direction, while the present study is considered to offer sufficient contributions through its systematic prompting analysis and meta-evaluation of automatic metrics.

Prompting. While this study adopts a systematic prompt strategy based on a taxonomy ranging from simple to complex prompts, many additional prompt variations are possible, such as more direct prompting strategies that explicitly incorporate singability-related phrases or approaches based on multi-agent systems. Moreover, further prompt optimization, including investigating how different instructions and prompt formulations influence system performance (Sakai et al., 2024; Suzuki et al., 2025), may further strengthen our method and analysis. In addition, we adopt conservative decoding parameter settings to ensure stable evaluation in this study. Accordingly, exploring more creative decoding strategies that better fit the characteristics of lyrics translation represents a promising research direction. Nevertheless, this work provides valuable insights into prompting design and evaluation methods for lyrics translation. These findings can serve as a basis for further validation and extension.

Data. In this study, all experiments are conducted using a dataset curated by the authors. As a result, the dataset is relatively small, comprising approximately 600 lyric sections. However, since our study focuses exclusively on zero-shot evaluation and the dataset covers a sufficient number of songs across multiple language directions, we consider it a reasonable and appropriate evaluation dataset for the scope of this work. In addition, due to the difficulty of recruiting native evaluators across a wide range of languages, we limit our experiments to three languages: Chinese, English, and Japanese. While extending the dataset to additional languages or conducting broader cross-lingual comparisons would be an ideal evaluation direction, the primary focus of this study is on prompting strategies. Therefore, larger-scale multilingual experimental settings are left for future work.

Ethical Considerations

License. We primarily collect our dataset from YouTube and Bilibili. The use of textual information such as lyrics from these platforms is not prohibited when used for research purposes, provided that the original content is not redistributed. For qualitative examples included in the paper, we confirm that, under the legal framework of the country in which our institution is located, disclosing a small portion of the content, such as a single lyric section without full disclosure, is permissible for research purposes and does not violate fair use provisions. Accordingly, such limited excerpts are included solely for illustrative analysis. In contrast,

redistributing full lyrics or audio content would pose a risk of copyright infringement and use beyond the scope of research, and is therefore explicitly avoided in this work. To support reproducibility, we instead make available experimental prompts, code, and derived annotations, to the extent that they do not conflict with licensing restrictions. This practice is consistent with established norms in prior research on lyrics. Based on these considerations, we confirm that the licensing aspects related to data usage in this study do not raise conflicts within the scope of our research.

Human Annotation. This study involves human evaluation and manual data construction. All annotators were provided with a clear explanation of the study purpose and procedures, and evaluations were conducted under a blind setting to prevent bias. As a result, the evaluation process does not involve arbitrary manipulation and ensures fairness and objectivity. In addition, all annotators completed agreements, including consent regarding the use and transfer of annotation results for research purposes. We ensure that the study complies with ethical standards for human subject research.

AI Tools. We used AI tools to assist with grammatical correction and translation. However, all research ideas, experimental design, and original writing were carried out by the authors. All AI-generated suggestions were carefully reviewed and verified by the authors prior to inclusion. Accordingly, full responsibility for the content of this paper rests with the authors, while we acknowledge the support provided by the AI tools used during manuscript preparation.

Acknowledgements

We thank the anonymous reviewers for their valuable comments and suggestions, which strengthened this work. We are also grateful to Dr. Kento Watanabe at the National Institute of Advanced Industrial Science and Technology (AIST) for insightful comments regarding the foundational knowledge of the background of this study. We sincerely thank him for sharing his knowledge and insights on lyrics translation, evaluation, and music information. We are pleased to report to him that this work has now been successfully completed and published.

This work has been supported by JSPS KAKENHI Grant Number 25K24369.

References

Woohyun Cho, Youngmin Kim, Sunghyun Lee, and Youngjae Yu. 2025. MAVL: A multilingual audio-video lyrics dataset for animated song translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13651–13679, Suzhou, China. Association for Computational Linguistics.

Yibo Hu, Erick Skorupa Parolin, Latifur Khan, Patrick Brandt, Javier Osorio, and Vito D'Orazio. 2024. Leveraging codebook knowledge with NLI and ChatGPT for zero-shot political relation classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 583–603, Bangkok, Thailand. Association for Computational Linguistics.

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

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Peter Low. 2003. Singable translations of songs. *Perspectives*, 11(2):87–103.

Peter Low. 2005. *The Pentathlon Approach to Translating Songs*, pages 185 – 212. Brill, Leiden, The Netherlands.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Katherine Hermann, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS) 2023*. Poster.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2024. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Poster.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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, Toronto, Canada. Association for Computational Linguistics.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ashmari Pramodya, Yusuke Sakai, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. Translating movie subtitles by large language models using movie-meta information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 315–330, Vienna, Austria. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Yusuke Sakai, Adam Nohejl, Jiangnan Hang, Hidetaka Kamigaito, and Taro Watanabe. 2024. Toward the evaluation of large language models considering score variance across instruction templates. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 499–529, Miami, Florida, US. Association for Computational Linguistics.

Wai Lei Song, Haoyun Xu, Derek F. Wong, Runzhe Zhan, Lidia S. Chao, and Shanshan Wang. 2023. Towards zero-shot multilingual poetry translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 324–335, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Barbora Štěpánková and Rudolf Rosa. 2025. Song lyrics adaptations: Computational interpretation of the pentathlon principle. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 117–128, Albuquerque, USA. Association for Computational Linguistics.

Toma Suzuki, Yusuke Sakai, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. Superfluous instruction: Vulnerabilities stemming from task-specific superficial expressions in instruction templates. In *Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM)*, pages 140–152, Vienna, Austria. Association for Computational Linguistics.

Shanshan Wang, Derek Wong, Jingming Yao, and Lidia Chao. 2024. What is the best way for ChatGPT to translate poetry? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14025–14043, Bangkok, Thailand. Association for Computational Linguistics.

Cheng Yang, Puli Chen, and Qingbao Huang. 2024. Can ChatGPT's performance be improved on verb metaphor detection tasks? bootstrapping and combining tacit knowledge. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1016–1027, Bangkok, Thailand. Association for Computational Linguistics.

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, Miami, Florida, USA. Association for Computational Linguistics.