Sing it, Narrate it: Quality Musical Lyrics Translation

Anonymous ACL submission

Abstract

Translating lyrics for musicals presents unique challenges due to the need to ensure high translation quality while adhering to singability requirements such as length and rhyme. Existing song translation approaches often prioritize these singability constraints at the expense of translation quality, which is crucial for musicals. This paper aims to enhance translation quality while maintaining key singability features. Our method consists of three main components. First, we create a dataset to train reward models for the automatic evaluation of translation quality. Second, to enhance both singability and translation quality, we implement a two-stage training process with filtering techniques. Finally, we introduce an inference-time optimization framework for translating entire songs. Extensive experiments, including both automatic and human evaluations, demonstrate significant improvements over baseline methods and validate the effectiveness of each component in our approach.

1 Introduction

Have you ever heard of Hamilton in Chinese, or Mamma Mia in Swedish (Åkerström, 2010)? Advancements in cultural globalization allow musicals to reach universal audiences, but language barriers still hinder full comprehension. Translating musicals into performing country's language enhances audience experience (Sorby et al., 2014) and expands commercial outreach (Andersson et al., 2008), as it allows audiences to enjoy theatrical elements without heavily relying on subtitles (Engel and Kissel, 2006; Sorby et al., 2014). However, musical translation is labor-intensive and time-consuming, requiring adjustments for musical framework, stage performance, and cultural references beyond mere verbatim translation (Sorby et al., 2014; Fei, 2014). To alleviate this burden, we aim to automatically translate musical lyrics from English to Chinese.

Song translation, a closely related field, requires aligning the translated text with the music to ensure the translated lyrics can be sung (Low, 2003; Franzon, 2005). However, musical translation requires an even higher standard of translation quality because lyrics play a crucial role in the story-telling of a musical (Kenrick, 2010; Carpi, 2020; Chan, 2017). To preserve the original narration, the translations must accurately convey the meaning and nuance of the source lyrics. This high fidelity ensures that the translated musical maintains its artistic integrity and allows the story to unfold as intended for the target audience. Thus, musical translation demands a rigorous approach to translation quality, focusing on maintaining the narrative function to create a faithful rendition of the original work.

To the best of our knowledge, there is no existing work on automatic musical translation, and existing works on automatic song translation (Guo et al., 2022; Ou et al., 2023; Li et al., 2023a) mainly focus on the alignment of text and music, sacrificing translation quality and often produce unnatural and inaccurate translations unsuitable for musicals, as shown in Figure 1. To distinguish our work from existing art, we focus on improving translation quality, which would contribute to maintaining the narrative function, while reasonably following singability constraints. We define translation quality using the well-established criteria for literature translation: fluency, accuracy, and literacy (Yan, 1898). Additionally, we consider the singability constraints of length and rhyme following previous works (Guo et al., 2022; Ou et al., 2023). Figure 1 shows our considered aspects, with examples demonstrating their significance.

To depict translation quality, we collect English-Chinese lyric pairs using large language models (LLMs), label them according to our scoring rubrics, and train reward models to provide evaluations that correlate with human scores. For singability constraints, we observe that LLMs struggle

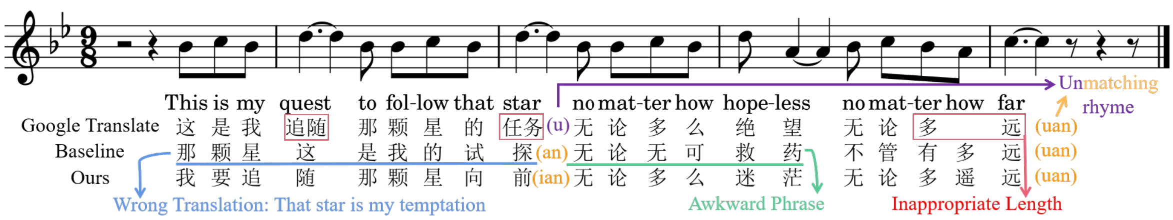

Figure 1: Aspects we considered include length, rhyme, and translation quality. The proper length of translated lyrics is the number of notes, and the end rhyme of each line (shown in parentheses) is better to have the same type (shown in the same color). Google translation fails to follow the length constraint and misaligns with music, as shown in red boxes, and its rhyme does not match either. Both baseline and our results meet length and rhyme constraints, but the baseline has inaccurate translations and inappropriate phrases, while our model generates higher-quality lyrics.

to adhere to them in a zero-shot manner. Thus, we perform two-stage translation model training to improve accuracy, balancing singability with translation quality using filtered high-quality data. Finally, to produce coherent translations for entire passages, we propose an inference-time optimization framework that leverages the output diversity of LLMs and a loss function designed to optimize paragraph-level overall quality. Extensive experiments demonstrate the effectiveness of our method's components, significantly outperforming the previous state-of-the-art approach.

To sum up, we make the following contributions: (1) We propose the task of musical translation, which requires a higher level of translation quality than song translation; (2) We create a dataset `MusicalTransEval` for scoring musical translation, which could serve as a valuable resource for future research; (3) We propose a two-stage translation model training approach that leverages reward models for data filtering and introduces a novel inference-time optimization framework, both aimed at improving translation quality while maintaining satisfactory singability performance.

2 Related Work

Translatology: Song and Musical Translation. In translatology, "Pentathlon Principle" (Low, 2003, 2005) is a well-known theory and guidance on general song translation (Franzon, 2008; Cheng, 2013; Stopar, 2016; Si-yang, 2017; Opperman et al., 2018; Sardiña, 2021; Pidhrushna, 2021; Ou et al., 2023), which proposes five criteria to consider: singability, rhyme, rhythm, sense, and naturalness, where the first three relates to music-text alignment and the rest refer to translation quality. However, this principle is not developed specifically for songs on the musical stage (Carpi, 2020).

The functional approach (Franzon, 2005) is more suitable for songs in musicals (Carpi, 2020), which emphasizes that the translated lyrics should replicate the function of the source text. In musicals, songs are "story-telling" elements (Kenrick, 2010), and the translated lyrics must carry out this role (Desblache, 2018; Åkerström, 2010; Sorby et al., 2014; Franzon, 2005). Thus a basic yet necessary constraint in musical translation is that lyrics must maintain the original narrative function, and thus should have high quality.

Automatic Song Translation. To our best knowledge, there are only three previous works on automatic song translation (Guo et al., 2022; Ou et al., 2023; Li et al., 2023a). Guo et al. (2022) mainly addresses the problem of aligning words' tones with the melody in the beam search phase, and Li et al. (2023a) focuses on aligning text to musical notes better. However, they both neglect the important rhyme constraint (Strangways, 1921). Ou et al. (2023) considers length, rhymes, and word boundaries, achieving decent results with prompting and the trick of reverse-order decoding. However, the translation quality is awkward and unsuitable for singing in musicals. To bridge this gap, we focus on generating high-quality translations under the two most important constraints for text-music alignment: length and rhyme.

LLM and Machine Translation. Recent years have witnessed the huge success of large language models (LLMs), including close-sourced GPT-4 (OpenAI, 2023), Kimichat, and open-sourced Llama-2 (Touvron et al., 2023). Recent works (Yang et al., 2023; Zhang et al., 2023; Zeng et al., 2024; Chen et al., 2023; Li et al., 2023b; Zhu et al., 2023) sought to enhance the machine translation capability using open-sourced LLMs, yet the improvements are limited. One challenge is balancing performance improvements during fine-tuning without significantly compromising the pre-trained

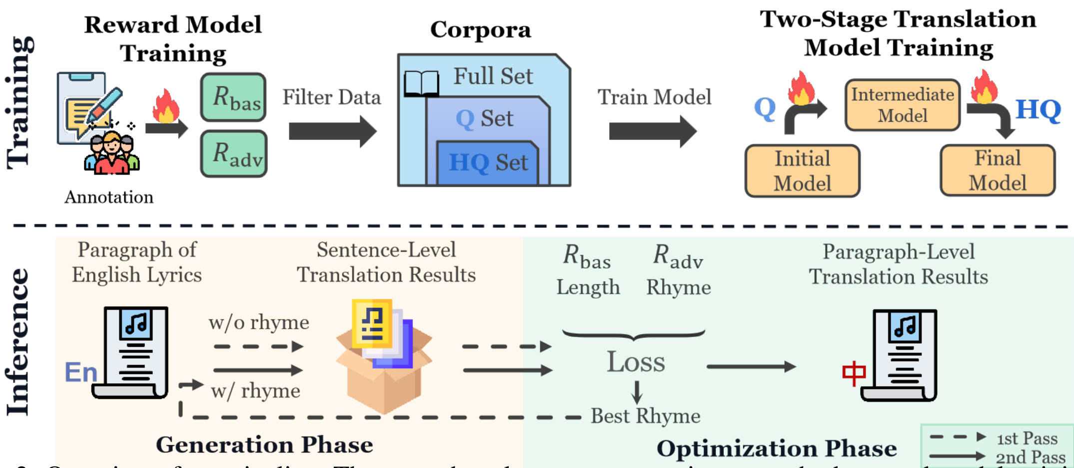

Figure 2: Overview of our pipeline. There are three key components in our method: reward model training (top left), translation model two-stage training (top right), and inference-time optimization framework (bottom). We use reward models to filter the whole corpora into a **Q**uality subset and a **H**igh-**Q**uality subset and train our generation model with the Q set and then with the HQ set. During inference, we generate plenty of sentence-level translations and derive paragraph-level translations by optimizing the loss function considering various aspects. We additionally give a 2nd pass with the same process but generate more sentence translations conditioned on the best rhyme.

model's knowledge. As Xu et al. (2024) pointed out, there is a diminished necessity for parallel data to fine-tune LLMs, and it is recommended to first train with monolingual data if the LLM does not have too much knowledge of the target language, and then fine-tune with a small amount of high-quality parallel data. Though our setting is slightly different, we similarly find it beneficial to fine-tune with high-quality parallel data.

3 Problem Formulation

We formulate the problem of musical translation as: Given a paragraph of English lyrics from a song, the task is to produce a Chinese translation that has high translation quality while adhering to singability constraints. By treating each paragraph independently, we can process an entire song.

To ensure **singability constraints**, we consider the following aspects. (1) *Length:* The number of syllables in the English lyrics and the number of characters in the Chinese lyrics should match the number of musical notes to ensure proper alignment. Since we lack direct access to sheet music but can easily obtain the English lyrics, we use the number of syllables in the English lyrics as the reference for alignment. (2) *Rhyme:* The translated sentences within each paragraph should maintain the same end rhyme as much as possible, particularly aligning with the end rhyme of the last sentence in each paragraph.

To evaluate **translation quality**, we focus on the following three aspects (Yan, 1898). (1) *Fluency*: The naturalness and readability of the translated lyrics in Chinese. (2) *Accuracy*: How well the translation conveys the same meaning as the original English lyrics. (3) *Literary quality*: The aesthetic appeal and literary merit of the translated lyrics. We further categorize fluency and accuracy as basic translation quality, while considering literary quality as advanced translation quality, to differentiate between mandatory and supplementary aspects. To enable machines to evaluate these aspects of translation quality, we train reward models using human annotations as learning data.

4 Method

Our method consists of three key components: a reward model trained to evaluate the quality of the translated language (Section 4.1), a translation model trained using a two-stage pipeline (Section 4.2), and an inference-time optimization framework that composes sentence-level results into paragraph-level output (Section 4.3). Figure 2 illustrates how these components work together.

4.1 Reward Model Training

To train our reward models to evaluate translations, we collect a dataset called `MusicalTransEval`, where each entry includes an original English line, a translated Chinese line, a paragraph as context, and three scores ranging from 1 to 4 that measure fluency, accuracy, and literacy of the translation respectively. The detailed scoring rubrics for each aspect are developed in collaboration with an expert in musical translation, and are shown in Appendix A. The English lines were extracted from musicals of diverse genres, ranging from fantasy, modern society, youth and family, history, and literature adaptation. The corresponding Chinese translations were generated by Kimichat using few-shot prompts. After 50 hours of annotation, we compiled a dataset with 3938 high-quality entries. For both basic and advanced translation quality, we train reward models using the dataset and refer to their evaluations as R_{bas} and R_{adv}, respectively.

To obtain a more balanced training dataset for R_{bas} and R_{adv}, we first apply mappings to handle categories that rarely appear. For R_{bas}, we map the score pairs of fluency and accuracy to a single integer score ranging from 1 to 4, resulting in 471, 322, 971, and 2174 entries, respectively. For R_{adv}, we map the scores for literacy to 2 or 3, obtaining 3104 and 834 data samples, respectively.

By utilizing data upsampling and downsampling techniques to further balance the training data, we obtained R_{bas} and R_{adv} with strong correlations with human judgments on a hidden balanced test set, which includes unseen musicals from the training period. The Pearson correlation (Pearson, 1895) of human scores with R_{bas} and R_{adv} are 0.649 and 0.532, signifying strong and moderate correlation. Besides, the precision and recall of the score 3 class R_{adv} are 0.95 and 0.49. The strong correlation of R_{bas} and high precision of R_{adv} make them quite reliable and valuable in our pipeline. More details of `MusicalTransEval` can be found in Appendix A and more training details are in Appendix B.

4.2 Two-Stage Translation Model Training

Large-scale training is essential to ensure the translation model generates results that accurately adhere to length and rhyme constraints, as discussed in Section 5.6. However, the same section also demonstrates that increasing the amount of training data does not always yield improvements in translation quality. This observation raises a pertinent question: how can we achieve high translation quality while maintaining satisfactory accuracy in terms of length and rhyme?

Due to the difficulty of collecting a large-scale musical dataset, we use the dataset provided by Ou et al. (2023), consisting of approximately 2.8M English-Chinese song lyrics sentence translations. To bridge the gap between normal and musical songs and improve dataset quality, we use our reward models to filter a high-quality subset of 1.75M and a higher-quality subset of 700K entries.

In the first training stage, we train the LLM with the large-scale high-quality dataset to primarily learn to follow length and rhyme constraints. In the second stage, we further refine translation quality by fine-tuning with the higher-quality dataset. In both training stages, we use the same prompt with length and rhyme constraints, ensuring that the constraints-following ability learned in the first stage is maintained in the second stage. Additional descriptions of the training dataset can be found in Appendix A and more translation model training details are in Appendix B.

4.3 Inference-Time Optimization Framework

Due to the inaccuracy of generating the whole paragraph at once, we let the translation model handle each sentence independently and then combine them using a novel optimization framework during inference. In particular, we design a proper paragraph-level loss function and optimize the overall loss by jointly considering all sentences.

In our setting, we consider length accuracy, rhyme score, and both basic and advanced translation quality. At the paragraph level, our overall loss $\mathcal{L}(\cdot)$ is defined for sentence-level translations $y_1, \ldots, y_n$ by incorporating all those aspects. Specifically, we define:

$$\mathcal{L}(y_1, \ldots, y_n) = \sum_i (\lambda_1[\text{Rhy}(y_i) \neq \text{Rhy}(y_n)]$$

$$+\lambda_2 D(\text{gt}_i, |y_i|) - \lambda_3 R_{\text{adv}}(y_i) - \lambda_4 R_{\text{bas}}(y_i)),$$

where:

$$D(y, x) = \begin{cases} \beta(x - y) & \text{if } y \leq x, \\ y - x & \text{if } y > x. \end{cases}$$

Here, $D(y, x)$ measures to which extent the translation length differs from the desired length, with an additional penalty β for translations that exceed the desired length, as this poses a greater challenge for singing. The two reward models R_{bas} and R_{adv} are introduced earlier. $\text{Rhy}(\cdot)$ specifies the rhyme type of the last character in a sentence, following the rhyme grouping rules from Xue (2002), a Chinese music translation book. Additional details of the loss function is in Appendix B.

Our goal is then to find a paragraph translation that minimizes the optimization objective. We select an appropriate temperature for the generation function and generate a diverse set of candidate translations for each sentence to ensure a broad

Method (Training Config.)	Rhyme	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
Ou et al. (2023)	yes	**0.977**	**0.96**	2.845	2.053	18.01	71.94
Ours VER.1 (1.75M)	yes	0.941	0.722	2.789	2.046	18.22	71.93
	no	0.854	-	2.92	2.053	17.15	71.61
Ours VER.2 (1.75M Q)	yes	0.914	0.687	2.971	2.056	18.32	72.87
	no	0.819	-	3.063	2.059	17.68	72.49
Ours VER.3 (1.75M Q + 700K HQ)	yes	0.923	0.703	**3.168**	**2.063**	18.80	**74.14**
	no	<u>0.874</u>	-	<u>3.248</u>	<u>2.068</u>	17.76	<u>73.78</u>

Table 1: Sentence-level results of the three versions of our method. In VER.1, we train the model with a 1.75M subset. In VER.2, we use a 1.75M **Q**uality subset. In VER.3, we use a 700K **H**igh-**Q**uality subset to fine-tune VER.2 model. Rhyme in the heading row means whether we use the rhyme constraint during inference, and the best results of the two cases are in **bold** (use) and <u>underline</u> (without use), respectively.

Method	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
Ou et al. (2023)	0.962	**0.95**	2.744	2.02	13.81	65.5
Ours VER.1	0.982	0.831	3.627	2.23	13.42	67.54
Ours VER.2	0.992	0.868	3.655	2.248	12.95	67.77
Ours VER.3	**0.99**	0.873	**3.76**	**2.248**	12.32	**69.43**

Table 2: The final whole-song translation results of three versions of our method. Compared with Table 1, our method includes the inference-time optimization framework here and can fully demonstrate our strength.

coverage of high-probability outputs in the generation space. This results in a vast number of possible combinations for $y_1, \ldots, y_n$. However, due to the structure of the optimization formula, we can solve it efficiently by first enumerating $\mathrm{Rhy}(y_n)$ for the last sentence, and then optimizing each sentence independently. It is worth mentioning that the flexibility of our proposed framework enables fine-grained control over additional singability constraints, which could be explored in future works.

After identifying the sentences $y_1, \ldots, y_n$ that minimize the loss function, we obtain an optimal rhyme. To ensure most sentences in a paragraph match the desired rhyme, we generate additional samples for each sentence with rhyme conditioning. The second pass is more focused and sample-efficient, as the desired rhyme is already fixed.

5 Experiments

In our experiments, we investigate the following research questions:

RQ 1 How well does our method perform in generating high-quality musical lyrics translations, as measured by automatic evaluation metrics?

RQ 2 How well do the generation results of our method align with human preference?

RQ 3 How does each component contribute to our performance improvements?

5.1 Experiment Configurations

Datasets. To evaluate musical translation performance, we additionally collect a dataset of English lyrics and quality Chinese translations from Cloud Music. This dataset includes 409 paragraphs and 1,741 lines from 56 popular songs of diverse musicals. More details can be found in Appendix A.3.

Models. For both the generation model and the reward model, we choose `Chinese-Alpaca-2-13B` (Cui et al., 2023) as our base model since it is pretrained with a large amount of Chinese corpora and has satisfying instruction-following ability.

Baselines. To the best of our knowledge, there are only three previous works on song translation, GagaST (Guo et al., 2022), Controllable Lyric Translation (Ou et al., 2023), and LTAG (Li et al., 2023a). Due to data acquisition difficulties of GagaST and LTAG, we have Ou et al. (2023) as our baseline. We train the baseline model directly using its released code.

Metrics. For automatic evaluation, we consider *length accuracy (LA)*, defined as the percentage of translated sentences whose length equals the desired length (we set it as the length of reference translation for sentence-level testing, and as the number of syllables of the English lyrics for paragraph-level testing), *rhyme score (RS)*, which is defined as the average percentage of sentences within each paragraph that exhibit identical end rhymes, *basic and advanced translation quality* R_{bas} *and* R_{adv} as defined in Section 4.1, statistic machine translation metric *BLEU* (Papineni et al., 2002), and model-based machine translation metric *COMET* (we use the `Unbabel/wmt22-comet-da` variant). (Rei et al., 2022). One caveat of BLEU is that it entirely depends on lexical form match and is sensitive to paraphrasing. On the other hand, COMET is robust and aligns much better with humans. COMET ranked 2nd in its alignment with humans among 20 metrics studied in Freitag et al. (2022), while BLEU only ranked 19th. Thus we mainly use COMET as the machine translation metric and report BLEU scores only for completeness.

5.2 Automatic Evaluations

The sentence-level performance of our generation models trained with several different recipes is reported in Table 1. In this experiment, we consider sentences in a paragraph as independent ones and set the desired length and rhyme according to our reference translation. We find that our dataset filtering strategy can largely improve translation quality by increasing all of R_{bas}, R_{adv}, and COMET. Also, after deleting the rhyme constraint in the prompt during inference time, generation results are still satisfactory even with slight improvements of R_{bas} and R_{adv}, though COMET slightly drops, partially due to the loss of length accuracy and therefore more misalignment with reference translation.

In this work, we focus more on the whole-musical translation results shown in Table 2, which again indicating that our training strategy is effective and both our two training stages can boost performance. Comparing our final results and the baseline result, it is evident that we have achieved significant improvements across the majority of metrics. The only metric that ours is not as good as the baseline is the rhyme score since Ou et al. (2023) uses its so-called reversed decoding technique to benefit rhyme following at the cost of language quality, but our rhyme score is already high enough for most applications, especially considering that even English lyrics in a paragraph does not guarantee the same rhyme. We thus answer **RQ 1** affirmatively: our method can indeed achieve much better translation quality while maintaining satisfactory singability performance.

5.3 Human Evaluations

We recruit 4 college students who are musical enthusiasts to do the human evaluation. We randomly sample 30 sentences and 12 paragraphs from our test set, let baseline and different versions of our model generate 120 sentences and 48 paragraphs, and ask another musical enthusiast to sing all generated results out. Subsequently, we let the evaluators assign scores on fluency, accuracy, literacy, and music-text alignment for sentence results, and overall translation quality and music-text alignment for paragraph results. We provide detailed scoring rubrics with examples and require the participants to adhere to our rules. This human evaluation can effectively demonstrate the subjective improvements of our methods over the baseline. See more details of human evaluation in Appendix D.

Method	Sentence-level				Paragraph-level	
	Fluency	Accuracy	Literacy	Alignment	Quality	Alignment
Ou et al. (2023)	2.88	2.53	2.37	2.48	2.08	2.92
Ours VER.1	3.09	2.6	2.45	2.69	2.31	2.75
Ours VER.2	3.25	2.64	2.54	2.6	2.27	**2.98**
Ours VER.3	**3.29**	**2.89**	**2.67**	**2.7**	**2.58**	2.96

Table 3: Human evaluation results. Our three versions correspond to those shown in Table 1, trained on different subsets: without filtering, with filtering, and with an additional second filtering.

The results shown in Table 3 are generally consistent with our automatic evaluations. The clear improvement of our VER.1 over the baseline and that of our VER.3 over the previous two versions demonstrate the effectiveness of our inference-time optimization and training dataset filtering. We thus answer **RQ 2** affirmatively: our method can align well with human preference and achieve better human evaluation scores.

We also note that although our rhyme accuracy is not as high as Ou et al. (2023), our singability scores in human evaluation are consistently higher than the baseline, indicating our rhyming accuracy is already good enough for human listeners. People might pay more attention to how we can hear the words clearly in the lyrics given music which could explain why we are seeing slightly improved results in text-music alignment.

5.4 Qualitative Results

In this section, we show a few representative qualitative results, with more results in Appendix C. For all Chinese translations, the translation errors and awkward phrases are underlined, and the excellent lyrics are underwaved.

Table 4 shows generation results of Ou et al. (2023), and our model. In our selected examples, the baseline has perfect rhyme, and only one failed length, but its translation quality is bad, with about one-third of incorrect or awkward phrases. For our model, the length is perfect, the rhyme accuracy is

Original lyrics	Ou et al. (2023)	Ours Ver.3
You are sixteen going on seventeen	你是十六个十七岁	你十六岁快要十七
Fellows will fall in line	伙伴们会结队	兄弟们排成排
Eager young lads and rogues and cads	渴望年少顽童和部队	年少轻狂的无赖痞子
Will offer you food and wine	献给你餐酒一杯	会为你提供美食
Sing once again with me,	再和我一起唱	和我再一起唱
our strange duet,	陌生的重唱	怪异对唱
my power over you,	我对你的力量	我对你的控制
grows stronger yet	更加茁壮	越来越强
Just because you find that life's not fair,	只因你发现生活不公平	只因为你发现生活不公
it doesn't mean that you just have to grin and bear it!	不代表只需要笑着忍痛	不等于只能强颜而忍耐
If you always take it on the chin and wear it	如果总是把它戴在你的头顶	如果总是硬着头皮强忍下来
Nothing will change.	不会变更	永不更改

Table 4: Qualitative results for ours, baseline, and Kimichat. Translational errors and awkward phrases are underlined. Excellent lyrics are underwaved.

Original lyrics	Ours Ver.1	Ours Ver.2	Ours Ver.3
Suddenly I'm flying company chaters	忽然间我飞去公司包机了	突然间我飞着公司的包机	突然间我正坐着包机飞往
Suddenly everything's high	突然什么都高涨	突然什么都高涨	突然一切都高涨
Suddenly there's nothing in between me and the sky	突然之间没有了我和天空相隔	突然之间隔着我和天空的天际	突然之间我和天空之间无屏障

Table 5: Qualitative results for our three versions corresponding to those shown in Table 1. They are trained on different subsets: without filtering, with filtering, and with an additional second filtering. Translational errors and awkward phrases are underlined. Excellent lyrics are underwaved.

Samples	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
1	0.862	0.385	3.061	2.074	12.79	67.94
80	**0.997**	0.862	**3.765**	**2.286**	12.32	68.84
40+40	0.99	**0.873**	3.76	2.248	12.32	**69.43**

Table 6: Comparison of no sampling, one-stage sampling, and our two-stage sampling strategy performance. 40+40 means the number of samples in two stages.

Reward	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
no	**1.0**	**0.94**	2.972	2.064	13.8	67.08
yes	0.99	0.873	**3.76**	**2.248**	12.32	**69.43**

Table 7: The comparison of whether there are reward model terms in the inference loss function, signified by Reward in the heading row.

satisfactory, and their translation results are fluent, correct, and sometimes impressive. Table 5 demonstrates the effectiveness of our training recipe. With further finetuning with high-quality data, the percentage of awkward phrases is reduced, and more excellent translations emerge.

5.5 Understanding the Contribution of Each Component

To answer **RQ 3**, we investigate the individual contribution of each component in our pipeline to the overall performance improvement.

Effectiveness of the optimization framework. Table 6 demonstrates the effectiveness of our optimization framework. If we forgo the optimization during inference and only rely on a single sampling step to obtain the final result, we observe significant drops across all metrics, particularly in the rhyme score. Compared to a simple one-pass strategy with equal computational resources (only using ensembling to fit a rhyme for a paragraph), incorporating a second stage enables us to achieve a better rhyme score by generating more rhyme-conditioned samples.

Impact of reward model terms in the inference loss. We additionally demonstrate that incorporating reward model terms in the inference-time loss is critical to the overall performance improvement. Under our best-performing configurations, removing the reward model terms from the optimization process results in a decrease of more than 2 points in the COMET score, as shown in Table 7. Compared to the one-sample setting in Table 6, the absence of reward model terms leads to a slightly larger drop in the COMET score, as the model attempts to optimize the rhyme score at the expense of translation quality.

Decomposing the sources of improvement. Compared to the work of Ou et al. (2023), while achieving comparable performance in terms of singability aspects, we analyze that the improvement in translation quality (approximated by the COMET score) can be primarily attributed to two factors. First, conducting dataset filtering using our trained reward models contributes to an improvement of approximately 2 points in the COMET score, as

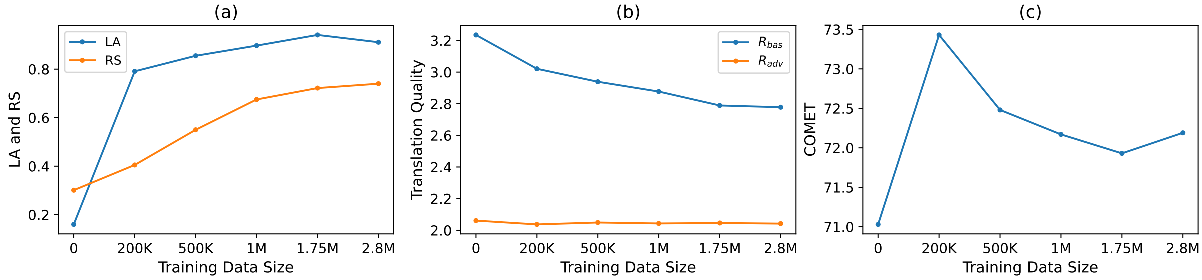

Figure 3: The changes of length accuracy, rhyme score, both basic and advanced translation quality, and COMET score if we change the training set scale.

Model	Trained	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
ours	no	0.844	0.574	3.731	2.159	12.49	68.1
ours	yes	**0.99**	**0.873**	3.76	2.248	12.32	69.43
Kimichat	no	0.944	0.669	**3.777**	**2.271**	15.98	**72**

Table 8: The comparison of closed-sourced Kimichat and both our untrained and trained model variants.

Samples	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
10+10	0.98	0.606	3.675	2.151	14.09	69.18
20+20	0.995	0.7	3.708	2.217	12.96	68.91
40+40	0.99	0.873	3.76	2.248	12.32	**69.43**
80+80	**1.0**	**0.906**	**3.777**	**2.269**	12.12	68.46

Table 9: Comparison of different numbers of samples in our framework, all using two sampling stages.

evidenced by Tables 1 and 2. Additionally, the inclusion of reward model terms in the loss function of our inference-time optimization framework provides a further improvement of 1.5 to 2 points in the COMET score, as shown in Tables 6 and 7.

5.6 Additional Analyses

Impact of training data scale. Figure 3 illustrates that increasing the scale of training data can help balance translation performance with length accuracy and rhyme score. Without training, the translation model struggles to adhere to length and rhyme constraints. As we increase the size of the training set, length and rhyme accuracy consistently improve, albeit at the cost of a slight drop in translation performance. This is expected, as our training helps the model follow the constraints but with imperfect translations, potentially diluting some of the pre-trained knowledge. To strike a balance, we use 1.75M data points to ensure high length and rhyme accuracy in the first training stage, and then employ high-quality filtered data to further refine translation quality in the second stage.

Potential benefits of larger closed-source LLMs. Our method is applicable to any pre-trained LLM, making it natural to explore the potential benefits of employing state-of-the-art closed-source LLMs for our task. We test our whole pipeline with Kimichat, given its strong understanding and expression capabilities in Chinese. We keep most components of our method unchanged except for translation model training due to inaccessibility. The results, shown in Table 8, indicate that Kimichat's zero-shot translation quality surpasses that of our fine-tuned `Chinese-Alpaca-2-13B`, but its length accuracy and rhyme score leave room for improvement. If we were able to apply our fine-tuning approach to the Kimichat model, we would likely observe better results, demonstrating the scalability of our method with respect to model size.

Effect of sample count in our framework. The number of samples used in our framework can be freely adjusted. As shown in Table 9, increasing the number of samples improves the rhyme score. We find that using 40 samples for both the first and second stages provides a good balance between performance and computational efficiency. This setting roughly takes 1 minute for each paragraph, which is quite acceptable in terms of the real-world musical lyrics translation application.

6 Conclusion

In conclusion, our work successfully balances translation quality and singability in musical lyrics translation. To solve this task, we leverage trained reward models, a two-stage translation model training approach, and an inference-time optimization framework. Our approach ensures that translated lyrics meet the criteria of fluency, accuracy, and literary quality while adhering to the critical constraints of length and rhyme. The substantial improvements over the baseline, as evidenced by both automatic metrics and human evaluations, demonstrate the efficacy of our method in delivering high-quality translations that retain the essence of musical expression. This work paves the way for future advancements in the field, advancing cross-cultural appreciation of musicals.

Limitations

Although the current version of our reward models can already achieve good results, there is room for further improvement by scaling the collected dataset and inviting more annotators to score sentence translations for less noise. We believe the power of reward models could be stronger if we can access more resources, making the results more impressive.

Besides, we are translating at the sentence level due to the difficulty of tackling various constraints and composing sentences into a paragraph. Yet in some cases, neighboring sentence translations are not that compatible. Thus to further improve translation quality, we believe it is a promising direction to explore how to directly translate a paragraph.

Also, in this work, we only consider two of the most critical singability aspects for simplicity. In future works, it is possible to consider more fine-grained singability constraints to make our compositions more professional.

Ethics Statement

This work addresses the task of musical translation, considering both translation quality and singability constraints. Potential risks include inaccurate translation results, which may lead to misunderstandings if used directly in certain scenarios.

The lyric data used in this research are sourced from the public Cloud Music platform and are used solely for research purposes. The models are obtained from public GitHub repositories. The dataset provided by Ou et al. (2023) is also used in accordance with its original intended purpose.

For human evaluations, we strictly adhere to the ACL Code of Ethics. Comprehensive details, including the recruitment process for evaluators and the instructions provided, are included in Appendix D. We collect evaluation scores without any personal information and ensure that the questionnaires do not contain offensive statements. Although our institute does not have an ethical review board or similar entity from which we can obtain approval, we have made every effort to follow the ethical guidelines set forth by ACL.

Regarding the use of AI assistants in our research, we primarily employed them for language polishing and refining the clarity of our writing. The main ideas, methodologies, and contributions presented in this paper are the result of our own work and intellectual efforts.

References

Johanna Åkerström. 2010. Translating song lyrics : A study of the translation of the three musicals by benny andersson and björn ulvaeus.

B. Andersson, B. Ulvaeus, J. Craymer, and P. Dodd. 2008. *Mamma Mia! How Can I Resist You?: The Inside Story of Mamma Mia! and the Songs of Abba*. Orion Publishing Group, Limited.

Beatrice Carpi. 2020. A multimodal model of analysis for the translation of songs from stage musicals. *Meta*, 65(2):420–439.

Candice Jing Harn Chan. 2017. *The" visible" translator: challenges and limitations in musical translation*. Ph.D. thesis.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023. Improving translation faithfulness of large language models via augmenting instructions.

Hui Tung Cheng. 2013. *Singable Translating: A Viewer-oriented Approach to Cantonese Translation of Disney Animated Musicals*. Ph.D. thesis, Chinese University of Hong Kong.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Lucile Desblache. 2018. Translation of music. *An encyclopedia of practical translation and interpreting*, pages 297–324.

L. Engel and H. Kissel. 2006. *Words with Music: Creating the Broadway Musical Libretto*. Applause Bks. Applause Theatre & Cinema Books.

Yuanhong Fei. 2014. 音乐剧翻译中的"信达雅". 上海戏剧.

Johan Franzon. 2005. *Musical Comedy Translation: Fidelity and Format in the Scandinavian My Fair Lady*, pages 263 – 297. Brill, Leiden, The Netherlands.

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

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

John Kenrick. 2010. *Musical theatre: a history*. Bloomsbury Publishing USA.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Chengxi Li, Kai Fan, Jiajun Bu, Boxing Chen, Zhongqiang Huang, and Zhi Yu. 2023a. 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, Singapore. Association for Computational Linguistics.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2023b. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions.

Low. 2005. The pentathlon approach to translating songs. *Song and Significance*.

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

OpenAI. 2023. Gpt-4 technical report.

Suezette Opperman, Marlie Van Rooyen, and Kobus Marais. 2018. An inter-semiotic approach to translation: Leonard cohen in afri-kaans. *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, 39(1):1–9.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Karl Pearson. 1895. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London Series I*, 58:240–242.

Olena Pidhrushna. 2021. Functional approach to songs in film translation: Challenges and compromises. In *SHS Web of Conferences*, volume 105. EDP Sciences.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Lucía Camardiel Sardiña. 2021. *The Translation of Disney Songs into Spanish: Differences Between the Peninsular Spanish and the Latin American Spanish Versions*. Ph.D. thesis, University of Hawai'i at Manoa.

Chen Si-yang. 2017. Practical strategies for devising singable song translations: A case study on wuhan university anthem translation. *Overseas English*.

S.L. Sorby, Hong Kong Baptist University. Department of English Language, and Literature. 2014. *Translating Western Musicals in Chinese: Texts, Networks, Consumers*. Hong Kong Baptist University.

Andrej Stopar. 2016. Mamma mia, a singable translation! *ELOPE: English Language Overseas Perspectives and Enquiries*, 13(1):141–159.

A. H. FOX Strangways. 1921. SONG-TRANSLATION. *Music and Letters*, II(3):211–224.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models.

Fan Xue. 2002. 歌曲翻译探索与实践.

Fu Yan. 1898. 译例言. 天演论译作.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. Tim: Teaching large language models to translate with comparison.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages.

A Dataset details

A.1 `MusicalTransEval` Dataset for Reward Model

For the `MusicalTransEval` dataset, we picked 11 musicals across various genres and spent about 20 hours extracting all the lyrics from their songs, breaking them down into paragraphs. The distribution of the musical genres is shown in Figure 4(a). Next, we used the Kimichat API to get initial translations for these paragraphs, tweaking our pipeline a bit: we kept the optimization but focused only on length and rhyme scores, as we did not have reward models yet. This gave us 15657 English lines. We then labeled 3938 of these lines in three different aspects, which took us another 30 hours. We divided the labeled data into training and test sets. Time and budget constraints meant we could not label everything, but what we did manage to label already gave us pretty good results.

Our labeling metrics for human labeling is shown in Figure 5, 6, 7. We let human label in three aspects: fluency, translation accuracy, and literary. Each aspect has 4 levels of scores, and we give instructions and examples for each level to ensure consistency among human scores. We have endeavored to ensure a scientific and rational scoring process, collaborating with domain experts to establish sound criteria that have gone through a few amendments during the preliminary labeling stage. Also, we ensure annotators have a good background of musicals and are familiar with the rubrics, thus trying our best to reduce bias in annotations.

A.2 Translation Model Training Dataset

As mentioned in Section 4.1, due to the difficulty of collecting a large-scale musical dataset, we use the dataset provided by Ou et al. (2023), which consists of approximately 2.8M song lyric sentence translations from English to Chinese for training. Although there is some gap between normal songs and musical songs, we bridge this gap and improve dataset quality by using our reward models to filter a high-quality subset of 1.75M and a higher-quality subset of 700K entries. The high-quality subset is obtained by selecting entries with a basic reward score $R_{\text{bas}} \geq 3$, while the higher-quality subset is derived by choosing entries with $R_{\text{bas}} = 4$. We observe that filtering the dataset using only the basic reward model already leads to improvements in the generated output. However, additionally employ-

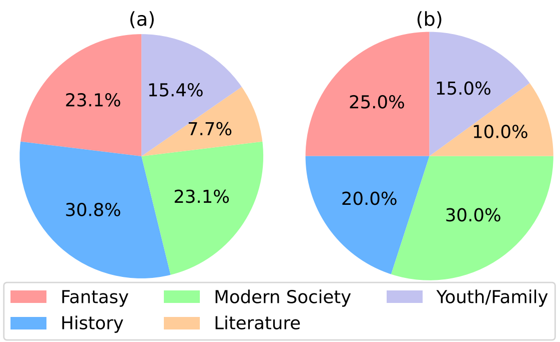

Figure 4: The distribution of musicals in `MusicalTransEval` dataset (a) and musical testing dataset (b).

ing the advanced reward model for filtering may result in overfitting, causing the generated lyrics to become overly flashy and less natural.

A.3 Musical Translation Test Dataset

We manually collect the lyrics from Cloud Music and split them into paragraphs. The length constraint is obtained by counting the syllables of the English lyrics using the Syllapy library. For testing BLEU and COMET scores, we collect the gold reference from human translations provided in Cloud Music. Our final musical dataset consists of 409 paragraphs and 1741 lines and mainly serves the purpose of testing performance. The musicals distribution is shown in Figure 4 (b).

B Implementation details

Reward Model Training Details. We use `Chinese-Alpaca-2-13B` for training reward models. See Table 10 for detailed prompts used for our two reward models.

For our basic translation quality reward model, there are 471, 322, 971, and 2174 data samples with scores from 1 to 4. We upsample class 2 with a ratio of 1.5, downsample class 3 with a probability of 0.7, and downsample class 4 with a probability of 0.5. After adjusting the training dataset, we train our model with 5 epochs. Data downsampling means we keep each data sample with some probability, and data upsampling with a ratio p means we first keep one copy of the dataset and then conduct data downsampling with probability $p - 1$ to derive additional data samples.

For our advanced translation quality reward model, there are 3104 samples with label 2 and 834 samples with label 3. We downsample class 2 with a probability of 0.4, upsample class 3 with a ratio of 1.5, and then train 5 epochs.

Translation Model Training Details. We also use `Chinese-Alpaca-2-13B` as the translation base model. See Table 10 for the prompts used for training. Both the two versions have the length constraint but one of them additionally has the rhyme constraint and is used in the second stage of the inference-time optimization framework. During translation model training, we mix the two prompts in the dataset so each data item appears twice (one with and the other without the rhyme constraint in the prompt).

We use 1 epoch for both training stages. Training on 1.75M data samples takes about 9 hours using 8 80GB A100 GPUs. The codebase is adopted from the DPO GitHub repository (Rafailov et al., 2023), which also supports supervised fine-tuning. We use the training batch size of 32 and keep all other hyper-parameters default choices in that repository.

Inference-time loss function Details. We explain details in the inference-time loss function here:

$$\mathcal{L}(y_1, \ldots, y_n) = \sum_i (\lambda_1 [\mathrm{Rhy}(y_i) \neq \mathrm{Rhy}(y_n)]$$

$$+\lambda_2 D(\mathrm{gt}_i, |y_i|) - \lambda_3 R_{\mathrm{adv}}(y_i) - \lambda_4 R_{\mathrm{bas}}(y_i)) \, .$$

The penalty coefficient in function $D(\cdot, \cdot)$ is set as $\beta = 2$. and the four hyperparameters are

$$\lambda_1 = 2, \lambda_2 = 3, \lambda_3 = 1, \lambda_4 = 1.$$

According to our rubrics, the translation basic quality is a compulsory requirement to ensure acceptable translation results, we thus only consider those with $R_{\mathrm{bas}} \geq 3$ to ensure translations are preferable. We may change to other hyperparameters to gain slightly better results, but in practice, this configuration can already achieve decent translation results.

Our pipeline with 40 + 40 samples runs within 8 hours on our musical test set and roughly takes 1 minute for each paragraph. In terms of real-world musical lyrics translation application, this speed is acceptable, thus during experiments we mainly focus on performance.

C More results

Table 12 showcases the qualitative effect of using reward models in the optimization framework. Without reward model terms, the translation quality significantly drops. Additional translation results are shown in Table 13.

We also put experiment results of different translation model inference configurations. Given the

Model	Prompt
Basic Reward Model	You are a translation grader. Given English lyrics and a corresponding Chinese translation, you need to give scores in the range of 1-4 (4 is the highest) considering both fluency and translation accuracy. Here are the metrics: Score 1: Not very fluent. There are inappropriate or awkward phrases or other big flaws. Score 2: Quite fluent, but there are serious translation mistakes that need correction. Score 3: Quite fluent, no big mistake in translation. But there are still small mistakes in phrasing or the translation of idioms. Score 4: Very fluent, no mistakes, and excellent translation. Note that a score of 4 means excellent and should be only given if you are absolutely sure the translated sentence is perfect. Any tiny mistake will make its score less than 4. Now, I will provide you with the English lyrics and the Chinese translation. You need to give me only one number and nothing else. For a comprehensive understanding, I will provide you the context: [paragraph]. The English lyrics is: [original lyrics]. The Chinese translation is: [translation]. The score is:
Advanced Reward Model	You are a translation grader. Given a Chinese translation of lyrics, you need to give scores in the range 1-4 (4 is the highest) for whether it looks like good lyrics. Criteria for scoring: Score 1: The translation does not resonate as good lyrics. Score 2: Acceptable as lyrics, but mundane and unremarkable. Score 3: Good fit for lyrics with some literary flair and aesthetic language. Score 4: Outstanding lyrical quality, inventive, expressive, and captivating. Reserve a score of 4 for truly impressive lyricism and be prudent when giving 4. Regular conversational phrases typically merit a score of 2. Now, I will provide you with the Chinese translation. You need to give me only one number and nothing else. The Chinese translation is: [translation]. The score is:
Translation Model w/o Rhyme	I will give you an English lyric and you need to translate it into Chinese with exactly [length] characters. Please only output the translated results and nothing more. The English lyrics are: [original lyrics]. Then the translation result is:
Translation Model w/ Rhyme	I will give you an English lyric and you need to translate it into Chinese with exactly [length] characters, where the ending rhyme type is [rhyme]. Please only output the translated results and nothing more. The English lyrics are: [original lyrics]. Then the translation result is:

Table 10: Prompts used for our two reward models and the translation model. For the translation model, we can only incorporate the length constraint or additionally add the rhyme constraint.

T	top-p	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
0.5	0.95	0.985	0.771	3.698	2.182	13.62	69.12
0.6	0.95	0.985	0.832	3.731	2.223	13.33	69.3
0.7	0.95	0.99	0.873	**3.76**	2.248	12.32	**69.43**
1	0.95	**1.0**	**0.901**	3.754	**2.325**	11.11	67.11
0.7	1	0.957	0.658	3.614	2.161	14.84	69.08

Table 11: Comparison of different sampling configurations (temperature and top-p probability).

importance of generating a large number of samples for ensembling, the sampling configuration plays a crucial role. Table 11 presents the results obtained by varying the temperature and top-p probability. With a lower temperature, the COMET score generally improves, as the outputs tend to have higher probabilities. However, this comes at the cost of reduced output diversity, resulting in a lower rhyme score. Conversely, increasing the temperature improves diversity but leads to a slight decrease in the COMET score. This trade-off between the COMET score and diversity is particularly pronounced in our constrained generation setting, where the space of acceptable solutions is often limited. We also investigate the effect of top-p sampling and find that it greatly enhances sample diversity, leading to improvements in both length accuracy and rhyme score, along with a slightly better COMET score. Based on these observations, we choose a temperature of $T = 0.7$ and top-$p = 0.95$, as this combination yields the best COMET score and high overall performance.

D Human Evaluation Details

We recruited 4 local college students who are musical enthusiasts from the college's musical club. We randomly sampled 30 sentences and 12 paragraphs from our test set, allowing the baseline and three versions of our model to generate 120 sentences and 48 paragraphs. We then asked another musical enthusiast to sing all the generated results. The evaluators assigned scores for fluency, accuracy, literacy, and music-text alignment for the sentence results, and overall translation quality and music-text alignment for the paragraph results. We provided detailed scoring rubrics with examples and required the participants to adhere to our rules. The English version of the instructions is shown in Figures 8, 9, and 10. Each annotator took 3 hours to complete the evaluations, and we compensated them with a reasonable price for university students.

To test the reliability of our human evaluations, we computed inter-rater agreement using intraclass coefficients (two-way mixed-effect, average measure model), following the practice of Ou et al. (2023). The results are as follows: 0.681 for sentence-level fluency, 0.727 for sentence-level accuracy, 0.546 for sentence-level literacy, 0.485 for sentence-level music-text alignment, 0.664 for paragraph-level overall translation quality, and 0.498 for paragraph-level music-text alignment. According to Koo and Li (2016), most of them fall into the "moderate reliability" range (0.5 to 0.75).

Original lyrics	Ours VER.3 with reward model	Ours VER.3 without reward model
Still strove, with his last ounce of courage,	还在竭尽全力地奋斗	拼了命的继续着奋搏
To reach the unreachable stars!	要飞到最远的星宿	去到那不曾到过的
Well, let that lonely feeling wash away	让那寂寞的感觉冲刷开	让那孤独感觉洗刷一洗
Maybe there's a reason to believe you'll be okay	也许有一些理由相信你会好起来	或许有理由相信你会过得很顺利
Cause when you don't feel strong enough to stand	当你感觉无力站起身来	因为你太懦弱无法站立
You can reach, reach out your hand	你可以伸出手来	伸出你的手可以

Table 12: The effects of using reward model terms in optimization pipeline. Translational errors and awkward phrases are underlined. Excellent lyrics are underwaved

Original lyrics	Translation results
I am I, Don Quixote, the Lord of La Mancha	我是我，堂吉诃德，拉曼查的领主
My destiny calls and I go	宿命呼唤，我随之去
And the wild winds of fortune, will carry me onward	命运的狂风，将我带向未知旅途
Oh whithersoever they blow	任凭风儿吹向何处
Whithersoever they blow, onward to glory I go	任凭风向何处吹，我直奔荣耀而去
Hamilton faces an endless uphill climb	汉密尔顿面临无尽上坡路
He has something to prove	他要证明什么
He has nothing to lose	他无后顾之忧
Hamilton's pace is relentless	汉密尔顿步履不停
He wastes no time	他毫不耽搁
What is it like in his shoes?	他脚下的路如何?
So let the sun come streaming in	就让阳光洒满房间
Cause you'll reach up and you'll rise again	因为你会奋起再登攀
Lift your head and look around	抬起头四处看看
You will be found	必被发现
you will be popular!	你会受到欢迎
You're gonna be popular!	你将会很有人气
I'll teach you the proper poise	姿势得体我来教
When you talk to boys	与男生谈笑
Little ways to flirt and flounce	小动作挑逗撒娇
To dream the impossible dream,	追求不可能的梦想
To fight the unbeatable foe,	挑战不可战胜之敌
To bear with unbearable sorrow,	承受那难以承受之痛
To run where the brave dare not go	勇闯无人敢去之地
I wrote my way out	我以笔自救
Wrote everything down far as I could see	写下所见所闻，尽我所能
I wrote my way out	我写下出路
I looked up and the town had its eyes on me	我抬头，全镇都在注视着我

Table 13: More qualitative results of our method, with Kimichat as the translation model. Excellent lyrics are underwaved

Evaluation Criteria

Sentence Completeness

[Only look at the Chinese, not the English]

1. Content is absurd, illogical, or incomprehensible at a glance

> Thou art base and debauched as can be
>
> 你艺术基地就有多颓废
>
> To love, pure and chaste, from afar,
>
> 爱，纯且贞，远远地
>
> Timid and shy and scared are you
>
> 又胆怯害怕你是谁

2. Mostly complete sentences, but with hard flaws (**unacceptable**), such as the use of very inappropriate words, lack of necessary components, serious ambiguity, or disordered syntax

> Your life, little girl, is an empty page,
>
> 女你的生活是空的一页 （首字"女"很不合适）
>
> Cuz for the first time in forever
>
> 第一次长久以来的 （语序混乱，应为"长久以来的第一次"）
>
> And I know they'll take you home
>
> 我知道，带你回家 （缺少主语，"他们"带你回家）

3. Mostly complete sentences, no hard flaws (**acceptable**), but may have awkward wording or minor ambiguities, slightly off from normal Chinese sentences

> For fate to turn the light on
>
> 命运点亮希望光 （"希望光"用词略显尴尬）
>
> When you're broken on the ground
>
> 你在地上摔碎了 （"摔碎"用词尴尬）
>
> But his voice filled my spirit with a strange, sweet sound
>
> 但那声音注入我灵魂，奇妙甜美嗯 （结尾的"嗯"比较尴尬）
>
> In sleep he sang to me
>
> 他梦里对我唱（有歧义，在谁的梦里？）
>
> For my own sanity, I've got to close the door
>
> 为保心神平衡，我需关门远离 （说不清哪里不对，但怪怪的）

4. Very smooth, easily understandable

> Cause when you don't feel strong enough to stand
>
> 当你感觉站不稳的时候
>
> Even when the dark comes crashing through
>
> 就算那黑暗突然袭来

Figure 5: Metrics for human labeling, page 1/3.

> Your life, little girl, is an empty page,
>
> 姑娘你的生活，如空白纸张

Translation Accuracy

[Only look at the translation's fidelity to the original meaning, regardless of sentence completion, consider context]

1. More than 50% of the translation is incorrect, or a few key parts (such as active/passive voice, verbs) are translated incorrectly or missing, **unacceptable**

   > Fellows will fall in line
   >
   > 兄弟长相厮守 （完全不对，应为“男人们会排队等待”）
   >
   > Tonight, we're gonna do ourselves justice,
   >
   > 今晚我们要做公正的自己 （关键部分不对，应为“今晚我们要为自己讨回公道”）
   >
   > I am sixteen going on seventeen
   >
   > 我是十六分继续十七分 （关键部分不对，应是“十六岁”、“十七岁”而非“十六分”、“十七分”）
   >
   > But now we're Ex-wives.
   >
   > 但现在，我们前妻。 （缺少谓语，我们“成为了”前妻）

2. Less than 50% of the translation is inaccurate, **barely acceptable** (allow for paraphrasing, allow for ignoring or changing a small amount of unimportant information)

   > Don’t know if I’m elated or gassy
   >
   > 不知我是欢喜还是气胀 （gassy在这里译为气胀不准确）
   >
   > And then I can go for a float
   >
   > 然后我能去漂浮了（“漂浮”不准确，应为游泳）

3. Basically accurate, but there is room for improvement, such as direct translation of English idioms without conveying the extended meaning, or adding a few small details would be better

   > Where in the world have you been hiding?
   >
   > 你在地球上藏哪儿了？ （俗语，翻译成“你到底藏在哪儿了”就可以）
   >
   > What is it like in his shoes?
   >
   > 穿他鞋，感觉如何？ （俗语in sb's shoes，翻译成“如果我是他”更好）
   >
   > Sven, the pressure is all on you
   >
   > 史文，压力都在肩头 （小瑕疵，应当是“压力都在你肩头”）
   >
   > Couldn't keep it in, heaven knows I've tried
   >
   > 实在忍不住，竭力试过了（keep it in“忍不住”稍有点奇怪）

4. Very accurate in meaning (allow for paraphrasing, allow for ignoring or changing a small amount of unimportant information)

   > I’ll be dancing through the night
   >
   > 我会跳舞到夜晚
   >
   > But you're dying to try

Figure 6: Metrics for human labeling, page 2/3.

> 但是你想尝试

Lyric Quality

[Only look at the Chinese, don't need to consider sentence completion]

1. Not like real lyrics

> That one man, scorned and covered with scars,
>
> 那一个人被伤疤抹掉

2. Suitable to be used as lyrics, and has a certain literary quality

> it doesn't mean that you just have to grin and bear it!
>
> 并不表示你只需要笑着忍痛
>
> In dreams he came
>
> 梦中他来
>
> When you're broken on the ground
>
> 当你破碎在原地

3. Suitable to be used as lyrics, and has a certain literary quality

> For the first time in forever
>
> 因为好久没在生命里
>
> That one man, scorned and covered with scars,
>
> 那一人，受辱满身伤痕
>
> In dreams he came
>
> 梦中降临
>
> To run where the brave dare not go;
>
> 勇闯，无畏者所不至
>
> the ground is falling backwards
>
> 地面倒退飞逝

4. Very suitable to be used as lyrics, creative, expressive, and **eye-catching**

> To run where the brave dare not go;
>
> 跋涉，无人敢行的路
>
> My destiny calls and I go
>
> 这命运召唤我启航！
>
> The sweet caress of twilight
>
> 暮光轻抚，甜如诗

Figure 7: Metrics for human labeling, page 3/3.

Human Evaluation Instructions

Our project use large models for musical translation. Given English lyrics, the model will automatically generate corresponding Chinese translations. We have used different models and methods to generate some results, and we ask you to score these results according to our established rules.

The test is divided into two parts. The first part scores individual sentences on translation quality and singability respectively. This part consists of 120 questions. The second part scores paragraphs, requiring both consideration of the lyric text and its coordination with music. This part has 48 paragraphs. We provide reference audio for lyrics involving music coordination.

Part One: Single Sentence Scoring

You will receive: a line of English lyrics, a Chinese translation, a paragraph containing this English lyric; a raw song snippet, and a reference audio of the lyrics being sung.

What you need to do: First, based solely on the text, score on fluency, translation accuracy, and literacy; then listen to the original song snippet and the translated audio to score the coordination of the translated lyrics with the music. Scoring standards are as follows.

Fluency (Consider only whether the Chinese text is coherent and fluent)

- 1 point: Not human language - content is absurd, illogical, or incomprehensible at a glance

  > 爱，纯且贞，远远地

- 2 points: Partially coherent, but with serious flaws (unacceptable), such as inappropriate vocabulary, missing necessary components, serious ambiguity, or disordered syntax

  > 第一次长久以来的 (disordered syntax, should be "长久以来的第一次")

- 3 points: Mostly coherent, without serious flaws (barely acceptable), but with awkward wording or minor ambiguities, slightly different from normal Chinese sentences

  > 命运点亮希望光 ("希望光" is an awkward term)

- 4 points: Very fluent, easy to understand the meaning

  > 当你感觉站不稳的时候

Accuracy (Combine the paragraph to judge whether the lyric translation is accurate)

- 1 point: More than 50% of the translation is wrong, or a small number of key parts (such as passive voice, verbs) are translated incorrectly or omitted, unacceptable

  > Fellows will fall in line
  >
  > 兄弟长相厮守 (completely wrong, should be "男人们会排队等待")

- 2 points: Less than 50% of the translation is imprecise, barely acceptable (allowing paraphrase, allowing the omission or change of a small amount of unimportant information)

Figure 8: Instructions for human evaluation, page 1/3.

> Don't know if I'm elated or gassy
> 不知我是欢喜还是气胀 ("gassy" does not translate correctly here)

- 3 points: Basically accurate, but there is room for improvement, such as direct translation of English idioms without conveying the extended meaning, or could add some small details to improve

> What is it like in his shoes?
>
> 穿他鞋，感觉如何？（The idiom "in sb's shoes" could be better translated as "如果我是他")

- 4 points: Very accurate in meaning (allowing paraphrase, allowing the omission or change of a small amount of unimportant information)

> To run where the brave dare not go
>
> 跋涉，无人敢行的路

Literacy (Consider only whether the Chinese text is suitable as a lyric)

- 1 point: Not like real lyrics

> 那一个人被伤疤抹掉

- 2 points: Can be used as lyrics, but plain and unremarkable, no highlights

> 并不表示你只需要笑着忍痛
>
> 当你破碎在原地

- 3 points: Suitable as lyrics, with a certain literary quality

> 因为好久没在生命里
>
> 那一人，受辱满身伤痕

- 4 points: Very suitable as lyrics, creative, expressive, and eye-catching

> 跋涉，无人敢行的路
>
> 这命运召唤我启航！

Single Sentence Evaluation of Lyric and Music Coordination

Mainly focus on three aspects:

- **Lyric word count**: Whether multiple words need to be crammed into one note, or one word corresponds to many notes? Generally, one note per word is the most suitable.
- **Pause**: Whether the pauses in the melody break up complete sentences/phrases? Ideally, the pauses in melody and semantics should coincide.
- **Misalign of tones and melody**: Is there a very serious reversal of words (hearing one word as another, such as "归来吧" heard as "鬼来吧")?

You don't need to consider translation accuracy here.

The audio examples for each score are in the file "Single Sentence Example.mp3".

Figure 9: Instructions for human evaluation, page 2/3.

- 1 point: The lyric word count is not perfect, it doesn't sound comfortable, there is room for improvement.

> For the first time in forever
>
> 在人生中第一次 (incorrect length)

- 2 points: The lyric word count is very suitable, but the pause is very inappropriate or there is a very serious reversal of words.

> is anybody waving back at me?
>
> 有没有人向我挥手回看 (There is a pause between "waving")

- 3 points: The lyric word count is very suitable, the pause is relatively suitable, the reversal of words is not very serious, but there are still strange-sounding places.

> To right the unrightable wrong.
>
> 解决对不对的事情("对不对"sounds strange, a bit of a reversal of words)

- 4 points: The lyric word count is very suitable, the pause is suitable, and the reversal of words is not serious.

> For the first time in forever
>
> 永远的第一次体验 (the coordination of lyrics and music is good)

Part Two: Whole Section Scoring

You will receive: a section of English lyrics, a Chinese translation, and a reference audio of the translated lyrics being sung.

What you need to do: For the whole section, score the lyric quality and its singability.

Whole Section Comprehensive Evaluation

Lyrics Quality:

- 1 point: Most of the lyrics are not human speech, or most of the lyrics deviate from the original meaning.
- 2 points: Most of the lyrics are human speech, but there are still a few awkward places (unacceptable), such as inappropriate wording or translation errors.
- 3 points: The lyrics are barely acceptable, but there are still flaws.
- 4 points: It's hard to tell it's a translation, it seems like the original Chinese lyrics.

Text-Music Alignment:

- 1 point: Very poor coordination of lyrics and music, such as many sentences with incorrect word counts, very un-rhyming in rhyming sections...
- 2 points: The overall coordination of lyrics and music is acceptable, but there are some awkward problems, such as unreasonable pauses, serious reversal of words...
- 3 points: There are no major problems with the coordination of lyrics and music, but there are still flaws.
- 4 points: It's hard to tell it's a translation, it seems like the original Chinese song.

Figure 10: Instructions for human evaluation, page 3/3.