MAVL: A Multilingual Audio-Video Lyrics Dataset for Animated Song Translation

Woohyun Cho[1], Youngmin Kim[1], Sunghyun Lee[1], Youngjae Yu[2*],
[1]Yonsei University [2]Seoul National University
k106419@yonsei.ac.kr youngjaeyu@yonsei.ac.kr

Abstract

Lyrics translation requires both accurate semantic transfer and preservation of musical rhythm, syllabic structure, and poetic style. In animated musicals, the challenge intensifies due to alignment with visual and auditory cues. We introduce Multilingual Audio-Video Lyrics Benchmark for Animated Song Translation (MAVL), the first multilingual, multimodal benchmark for singable lyrics translation. By integrating text, audio, and video, MAVL enables richer and more expressive translations than text-only approaches. Building on this, we propose Syllable-Constrained Audio-Video LLM with Chain-of-Thought (`SylAVL-CoT`), which leverages audio-video cues and enforces syllabic constraints to produce natural-sounding lyrics. Experimental results demonstrate that `SylAVL-CoT` significantly outperforms text-based models in singability and contextual accuracy, emphasizing the value of multimodal, multilingual approaches for lyrics translation.

1 Introduction

Lyric translation, a specialized task, prioritizes "singability"—fitting lyrics to melody. This often requires beyond-literal translation to preserve both musicality and meaning, making it significantly more complex than standard text translation.

However, text-based lyric translation has inherent limitations, especially in musical animations. For example, consider the lyric "And there's a butterfly" from the song "Get Back Up Again" in the movie Trolls. A text-only translation, such as one provided by Google Translate, might yield "그리고 나비가 있습니다" (Geu-ri-go na-bi-ga it-seum-ni-da), which literally means "And there's a butterfly." While this conveys the basic presence of a butterfly, it lacks the dynamic action depicted visually and offers poor singability. In contrast, by incorporating audio and video context, a multimodal system like `SylAVL-CoT` can produce a translation such as "나비가 날아와" (Na-bi-ga na-ra-wa), meaning "A butterfly comes flying". This version, as shown in Figure 1, is more vivid, aligns with the on-screen motion, and demonstrates superior singability, naturalness, and human-likeness compared to the text-only approach. Appendix G details `SylAVL-CoT`'s multimodal reasoning for context-aware translations, essential for musical and cinematic cohesion through rhythm and visual storytelling. Furthermore, cross-lingual syllabic and rhythmic differences necessitate adaptations beyond literal translation, incorporating musical elements for naturalness and markedly increasing complexity over standard text translation.

However, despite this importance, previous studies rely on text-based or text-and-score-based approaches, limited by musical constraints and specific languages (Guo et al., 2022; Kim et al., 2024; Li et al., 2023; Ye et al., 2024). To address these limitations, we introduce the MAVL, **M**ultilingual **A**udio-**V**ideo **L**yrics Benchmark for Animated Song Translation. MAVL is a novel benchmark for multilingual, multimodal lyric translation in animated musicals (Figure 1), featuring aligned lyrics (English, Spanish, French, Korean, and Japanese) with audio-video data. This allows models to integrate textual, auditory, and visual information for more contextually and emotionally resonant translations.

Leveraging MAVL effectively requires models that jointly process text, audio, and video while maintaining linguistic and musical coherence. To address current limitations in handling such multimodal data, we propose **Syl**lable constrained **A**udio-**V**ideo **LLM** with **C**hain **o**f **T**hought (`SylAVL-CoT`). `SylAVL-CoT` enhances standard reasoning by incorporating audio and video cues, enabling better integration of contextual information

*Corresponding Author
Code: k1064190/MAVL
Dataset: NoeName/MAVL

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 13640–13668
November 4-9, 2025 ©2025 Association for Computational Linguistics

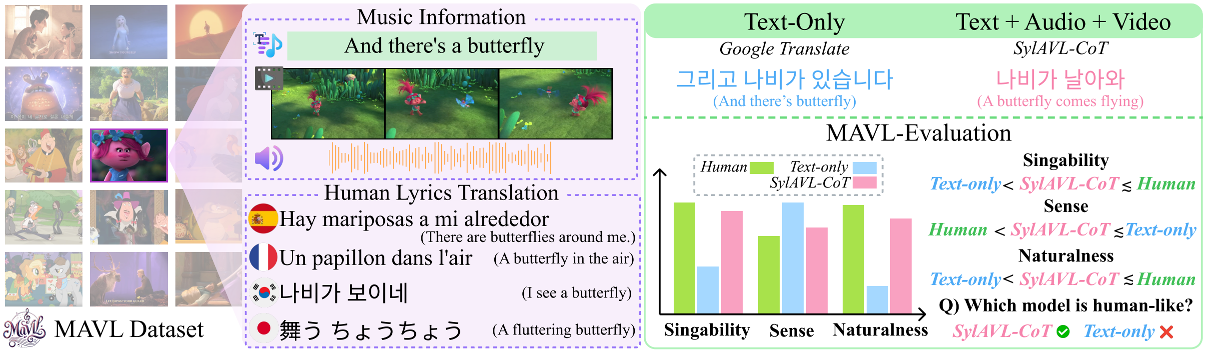

Figure 1: **Overview of MAVL Benchmark.** This lyric example is part of the OST "Get Back up Again" from "Troll", produced by Disney Corporation. the left illustrates MAVL Dataset components (music, multilingual human lyrics). The right compares translation outputs, showing our audio-visual `SylAVL-CoT` produces more vivid and human-like results than text-only models.

across modalities for more accurate and musically coherent lyric translations.

Furthermore, to systematically evaluate multilingual, multimodal lyric translation models leveraging MAVL, we propose evaluation metrics for more human-like results. Using these metrics, we analyze MAVL, quantitatively and qualitatively evaluate our approach against existing models, and conduct ablation studies demonstrating the necessity of multimodal input and chain-of-thought stages for effective lyric translation.

In summary, our main contributions are:

- We introduce **MAVL**, a multilingual, multimodal dataset and benchmark for multimodal lyric translation in animated musicals, consisting of aligned text, audio, and video data across five languages.
- We establish evaluation metrics for multilingual lyrics translation.
- We propose a `SylAVL-CoT` for lyric translation that enhances standard reasoning by incorporating audio and video cues.

2 Related works

Lyrics Translation Challenges and Strategies in Translation Studies. Translating lyrics has long been recognized as a specialized domain in translation studies, as it must balance semantic equivalence, poetic structure, and musical requirements (Franzon, 2008; Low, 2003, 2005). Early frameworks propose strategies ranging from literal translation to complete adaptation, guided by the "Pentathlon Principle" (singability, sense, naturalness, rhythm, and rhyme). Research on musicals and Disney soundtracks highlights deliberate manipulation of rhyme schemes and syllable counts to maintain musical flow, alongside cultural shifts for humor and emotional nuance (Leni and Pattiwael, 2019; Susam-Sarajeva, 2008). In audio-visual translation (AVT), filmic elements such as camera angles, music tempo, and background music, especially in musicals and animations is important (Baños Piñero and Chaume, 2009; Carpi, 2020; Taylor, 2016; Pidhrushna, 2021). Certain scenes rely heavily on imagery or character expressions to convey emotional subtext (Supardi and Putri, 2018), making strict fidelity to source lyrics potentially mismatched with the visual narrative.

Lyrics Machine Translation. Building on the principle that lyric translation prioritizes "singability," recent work has advanced the field through dataset creation, joint learning of melodic and textual features, and development of specialized evaluation metrics. For instance, (Guo et al., 2022) incorporated tonal constraints for translating lyrics into Mandarin, while (Ou et al., 2023) experimented with integrating melody-length and phonetic constraints into translation systems. Also, (Kim et al., 2023, 2024) developed a K-pop lyric translation dataset and trained a model. Further advancements include (Li et al., 2023), who presented a method for jointly learning melody and lyric semantics, and (Ye et al., 2024), which demonstrated a system that successfully fuses semantic fidelity with musical coherence.

Multimodal Chain-of-Thought Reasoning. Multimodal Chain-of-Thought (CoT) reasoning (Zhang et al., 2024; Ma et al., 2025) extends traditional CoT prompting (Wei et al., 2022) by incorporating vision modality to enhance complex reasoning

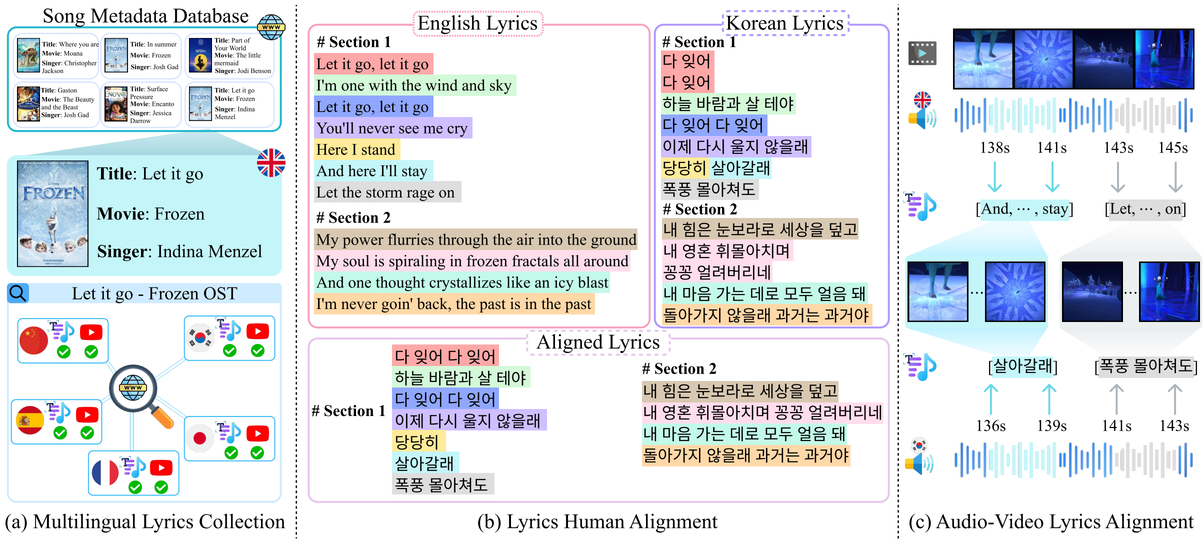

Figure 2: **MAVL dataset collection pipeline.** (b) visualizes the lyric alignment process, where each color corresponds to English and Korean, respectively. This example lyrics and images are part of the OST "Let it go" from Frozen, produced by Disney Corporation. For more details, refer to Section 3.1

tasks (Wang et al., 2024; Xu et al., 2024). Moreover, CoT prompting has been utilized to generalize reasoning across diverse languages and specialized knowledge domains (Hu et al., 2024b).

Beyond these advancements, CoT prompting has been utilized in Multimodal machine translation tasks (Barrault et al., 2018) to improve translation quality (Rajpoot et al., 2024; He et al., 2024) and has also been applied to machine translation evaluation (Qian et al., 2024). Building on this, multimodal CoT techniques have also been explored in speech translation (Du et al., 2024; Hu et al., 2024a).

3 Multilingual Audio-Video for Lyrics Translation Benchmark

We introduce MAVL, the Multilingual Audio-Video Lyrics Benchmark for Animated Song Translation, designed to evaluate lyric translation models integrating text, audio, and video, ensuring linguistic and musical coherence. MAVL comprises three key components: 1) a dataset of aligned lyrics, audio, and video (Section 3.1), 2) a syllable-constrained audio-video LLM with the chain-of-thought called `SylAVL-CoT` (Section 3.2), and 3) evaluation metrics (Section 3.3).

3.1 Multilingual Audio-Video for Lyrics Translation Dataset

This section details our MAVL dataset collection pipeline, illustrated in Figure 2. The dataset contains lyrics and corresponding audio-video data for

Datasets	Lang	Songs	Modality	Singablilty	Available
(Li et al., 2023)	En, Zh	79	Text, Score	✓	✗
(Kim et al., 2024)	En, Ko	1,000	Text	✓	✓
(Ye et al., 2024)	En, Zh	11*	Text	✗	✗
(Ou et al., 2023)	En, Zh	5,341	Text	✗	✓
MAVL	En, Fr, Es, Ko, Ja	228	Text, Video, Audio	✓	✓

Table 1: **Comparison of lyrics translation parallel datasets.** "*" indicates the number of musicals. The number of songs per musical is not specified in the respective paper.

228 songs across five languages (English, Spanish, French, Japanese, and Korean), as summarized in Table 2. To our knowledge, MAVL is the first dataset to support multilingual lyric translation across three modalities. Further details are in Appendix D.

Multilingual Lyrics Collection. We first gathered metadata (song titles, artists) for English animated film music from `last.fm`. Original English lyrics were then retrieved from `genius` and manually verified against the songs. Based on these verified English song titles, we proceeded to collect non-English lyrics corresponding to official dubbed versions of the original English songs in four languages: Spanish, French, Korean, and Japanese. We utilized platforms such as lyricstranslate[1]. Searching this platform by the original English song title typically yields a list of that song's lyrics in various languages. These listed versions often represent the official, singable dubbed interpretations, distinct from literal, non-singable translations that might be

[1]https://lyricstranslate.com/

Language	# Songs	# Video	# Sections	# Lines
English	228	228	1,923	6,623
Spanish	201	181	1,595	5,739
French	158	143	1,421	4,821
Japanese	138	114	1,264	4,280
Korean	133	117	1,138	3,974

Table 2: **Statistics of the MAVL Benchmark Dataset.** "# Section" refers to sections of the lyrics, while "Lines" denotes the individual lines within those sections. The number of videos is equal to the number of audio.

found if browsing general "translation" categories. Alongside these candidate lyrics, we gathered the corresponding audio and video.

Lyrics Human Alignment. The human alignment process was crucial not only for segmenting lyrics but also for rigorously verifying their authenticity as official, singable dubbed versions. This involved simultaneously cross-referencing the original English audio/video with the non-English candidate lyrics and their corresponding official dubbed audio/video. During this stage, candidate non-English lyrics were critically evaluated: if a set of lyrics could not be confirmed against an official audio-visual release, or if they did not accurately match the sung content in the verified official dub, that specific language version was excluded from our dataset. This process ensured that only verified, officially dubbed lyrics were retained. During alignment, non-singable dialogue or overlapping lyrics (prioritizing main melody) were also excluded.

Audio-Video-Lyrics Alignment. To align audio, video, and lyrics, we utilized `stable-ts`[2], a Whisper model (Radford et al., 2022)-based tool, to generate stable timestamps. This allowed us to determine the start and end times of each lyric line and segment the corresponding audio and video, extracting synchronized audio and video information aligned with the lyrics. Detailed alignment methodology, including our ensemble approach and quality assurance process, is provided in Appendix D.2.

3.2 Syllable-Constrained Audio-Video LLM with the Chain of Thought

In this section, we introduce Syllable-Constrained Audio-Video LLM with Chain of Thought (**`SylAVL-CoT`**), a multimodal approach designed to enhance lyrics translation by integrating audio, video, and text while maintaining rhythmic and semantic coherence. Existing Multilingual Large Language Models (MLLMs) struggle to integrate audio, video, and text, and the lack of aligned multilingual audio-visual datasets makes fine-tuning impractical. Additionally, they fail to maintain syllabic alignment, leading to unnatural phrasing and disrupted singability in translated lyrics.

[2]https://github.com/jianfch/stable-ts

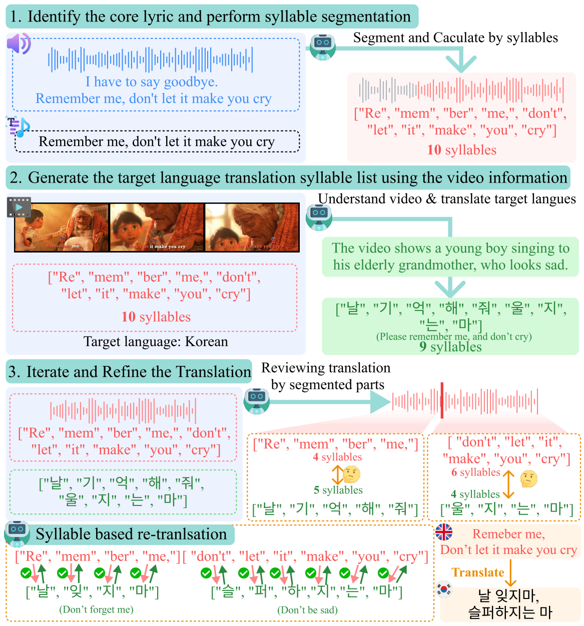

Figure 3: **`SylAVL-CoT` pipeline for lyrics translation.** This three-step process segments syllables utilizing audio, translates using video context, and iteratively refines the output to match original syllable counts.

To address this, we leverage Gemini 2.0 Flash (DeepMind, 2024) and adopt Chain-of-Thought (CoT) reasoning (Wei et al., 2022) to incorporate multimodal cues and enforce syllable constraints. Specifically, we implement syllable-aware alignment, where the model dynamically adjusts phrasing to match the original syllable count and rhythmic structure. The syllable constraint is applied during inference by providing the model with the exact syllable count of the source lyrics via prompting, instructing it to generate translations that match this count as closely as possible (see Appendix F for detailed prompts). Our **`SylAVL-CoT`** follows three-step process, which is illustrated in Figure 3. Examples of the model's detailed reasoning trajectories for lyrics translation are provided in Appendix G.

Identify the Core Lyric and Perform Syllable Segmentation. We begin by supplying the model with a specific segment of the original lyric text, accompanied by an audio snippet. The model's task is to locate precisely which part of the audio is relevant to the text. Next, the system carefully segments the lyric into syllables based on audible breaks in the audio—an essential first step for maintaining the original rhythm and singability. This initial segmentation guides subsequent steps, providing a structural template for the translation.

Generate the Target-Language Translation Syllable List, Utilizing Video Context. In the second stage, the pipeline processes visual cues (e.g., thematic elements, animation style, cultural context) from the video to refine the translation. The model aims to capture not only the literal meaning of the lyrics but also subtleties related to imagery, cultural nuances, and artistic style. During this step, the model strives to preserve the original syllable count to maintain or approximate the musical flow.

Iterate and Refine the Translation. Preserving syllable count, natural flow, and rhythmic fidelity can be challenging—especially when translating between languages with differing grammatical structures and phonetic inventories. To handle this complexity, the model iteratively refines the translated text by paraphrasing or reordering words until it achieves a final output that is both culturally appropriate and linguistically coherent. The model checks the number of syllables against the target and, if needed, continues to adjust the translation or segmentation.

3.3 Evaluation Metrics

We evaluate lyric translation quality across three principles. *Singability* evaluates whether the translated lyrics are suitable for singing, *Sense* is about accurately conveying the meaning and message of the original lyrics, and *Naturalness* evaluates whether the translated lyrics sound natural and conversational in the target language. Details are in Appendix A.

Syllable Error. Existing lyric translation evaluation metrics (Kim et al., 2023) primarily rely on the syllable count of the original language, failing to capture the subtle nuances of multilingual lyric translation, which can negatively impact *Singability*. In particular, simply comparing syllable counts without considering phonological and rhythmic differences across languages risks compromising the naturalness of translated lyrics. Therefore, it is essential to incorporate dubbed lyrics that adapt syllable counts.

Syllable error (SE) measures how well the syllable count aligns with the original English lyrics (c_{en}) and how closely it matches the dubbed lyrics, which reflect the linguistic characteristics of the target language (c_{dub}). For the purpose of explanation, we will denote both c_{en} and c_{dub} as c. Given the syllable counts, c and syllable counts of machine-translated text, c_{pred}, syllable error (SE) is calculated as follows:

$$SE = \begin{cases} c - c_{pred}, & \text{if } c \geq c_{pred} \\ \beta(c_{pred} - c), & \text{if } c < c_{pred} \end{cases} \tag{1}$$

where $\beta \geq 1$ is a penalty factor for exceeding the reference syllable count. We set the additional penalty $\beta = 2.0$ in our experiments as suggested by (Ye et al., 2024), to penalize exceeding the syllable count more heavily, as it can be more detrimental to *Singability*. In addition, we also employ the Syllable Count Distance (SCD) error rate proposed in (Kim et al., 2024). SCD is defined as:

$$SCD = \frac{1}{2}\left(\frac{|c - c_{pred}|}{c} + \frac{|c - c_{pred}|}{c_{pred}}\right) \tag{2}$$

Finally, the error rate is defined as the proportion of lines whose predicted syllable counts do not match the reference (original or dubbed) syllable counts.

Semantic scores. Accurately conveying the semantic meaning of lyrics is crucial in lyric translation. While previous studies (Li et al., 2023; Ou et al., 2023) primarily relied on word-overlap-based evaluation metrics such as BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) to assess semantic similarity, these approaches have clear limitations when translating creative content that demands both fluency and inventiveness.

To address these limits, we utilize cosine similarity measures using multilingual embeddings from pre-trained sentence embedding models, namely mpnet-base-v2 (Song et al., 2020), as follows:

$$sim_M = \cos(e^M_{gt}, e^M_{pred}), \tag{3}$$

where, M represents the model (mpnet-base-v2), e_{gt} and e_{pred} denotes the model's embeddings of ground truth and machine-translated lyrics, and $\cos(\cdot,\cdot)$ refers to the cosine similarity function. We treat both the original English lyrics and the dubbed lyrics as ground truths. We also performed experiments using other multilingual embedding models in Appendix C.

Phonetic Distance. We evaluate *Naturalness* by assessing pronunciation similarity. To do this, we convert the lyrics into International Phonetic Alphabet (IPA) (Laver, 1994) and measure the phonetic

	Model	Syllable Error (SE) ↓				Syllable Distance↓				Error Rate↓			
		ES	FR	KO	JA	ES	FR	KO	JA	ES	FR	KO	JA
		English Lyrics ⟷ Translated Lyrics											
	Human Expert	1.780	1.031	1.052	1.716	0.115	0.098	0.141	0.096	0.587	0.455	0.353	0.607
w/o SC	Google Translate	7.784	4.140	12.226	20.924	0.411	0.248	0.619	0.951	0.926	0.818	0.943	0.975
w/o SC	mBART-50	25.790	4.397	13.036	15.617	1.443	0.268	0.650	0.741	0.953	0.831	0.954	0.956
w/o SC	Qwen2.5-72B	8.063	4.132	9.870	15.654	0.426	0.249	0.515	0.750	0.939	0.816	0.939	0.973
w/o SC	GPT-4o	8.100	4.158	10.302	16.711	0.420	0.245	0.524	0.780	0.950	0.824	0.947	0.984
w/o SC	Gemini	8.718	4.660	10.819	16.919	0.456	0.276	0.557	0.798	0.948	0.829	0.944	0.977
w/ SC	Qwen2.5-72B	6.512	2.729	6.502	10.061	0.354	0.183	0.361	0.517	0.927	0.754	0.911	0.962
w/ SC	GPT-4o	3.164	1.442	3.084	7.221	0.185	0.138	0.182	0.378	0.795	0.676	0.935	0.935
w/ SC	Gemini	3.585	1.384	3.039	6.257	0.207	0.104	0.190	0.389	0.872	0.604	0.848	0.952
w/ SC	SylAVL-CoT (Ours)	0.966	**0.902**	0.695	2.572	0.063	0.089	0.049	0.155	0.352	0.510	0.237	0.611
w/ SC	SylAVL-CoT (Gemini 2.5)	**0.918**	0.909	**0.660**	**2.507**	**0.059**	**0.086**	**0.046**	**0.151**	**0.334**	**0.498**	**0.225**	**0.596**
		Dubbed Lyrics ⟷ Translated Lyrics											
w/o SC	Google Translate	6.513	4.157	11.848	20.398	0.340	0.272	0.624	0.928	0.886	0.806	0.944	0.967
w/o SC	mBART-50	24.481	4.424	12.651	15.182	0.126	0.292	0.655	0.729	0.929	0.837	0.950	0.945
w/o SC	Qwen2.5-72B	6.745	4.158	9.518	15.165	0.351	0.275	0.523	0.928	0.900	0.807	0.936	0.964
w/o SC	GPT-4o	6.717	4.237	9.907	16.227	0.341	0.289	0.538	0.768	0.894	0.814	0.931	0.974
w/o SC	Gemini	7.347	4.658	10.436	16.364	0.377	0.299	0.560	0.772	0.908	0.831	0.936	0.964
w/ SC	Qwen2.5-72B	5.185	2.893	6.229	9.703	0.279	0.222	0.370	0.509	0.876	0.758	0.898	0.942
w/ SC	GPT-4o	2.403	1.723	3.067	6.940	0.162	0.183	0.212	0.338	0.742	0.709	0.906	0.906
w/ SC	Gemini	2.610	1.703	2.941	5.996	0.163	0.150	0.205	0.382	0.759	0.667	0.813	0.917
w/ SC	SylAVL-CoT (Ours)	1.349	1.363	1.261	3.107	0.125	0.149	0.122	0.223	0.634	0.631	0.472	0.759
w/ SC	SylAVL-CoT (Gemini 2.5)	**1.299**	**1.269**	**1.200**	**2.979**	**0.120**	**0.139**	**0.116**	**0.214**	**0.611**	**0.603**	**0.449**	**0.728**

Table 3: **Comparison with other models: Syllable errors.** "SC" means syllable constraint. The syllable constraint is a condition applied when using LLM models. If the syllable constraint is not applied, the model is prompted to perform a simple translation. Conversely, when the syllable constraint is applied, the syllable count from the original lyrics is provided as a condition. **Best** and second are highlighted.

	Model	MPNet↑			
		ES	FR	KO	JA
		English Lyrics ⟷ Translated			
	Human Expert	0.639	0.613	0.575	0.537
w/o SC	Google Translate	**0.905**	**0.903**	**0.857**	**0.857**
w/o SC	mBART-50	0.766	0.890	0.850	0.842
w/o SC	Qwen2.5-72B	0.900	0.899	0.841	0.836
w/o SC	GPT-4o	0.899	0.897	0.849	0.842
w/o SC	Gemini	0.893	0.891	0.843	0.834
w/ SC	Qwen2.5-72B	0.883	0.878	0.823	0.810
w/ SC	GPT-4o	0.830	0.842	0.793	0.782
w/ SC	Gemini	0.781	0.790	0.754	0.733
w/ SC	SylAVL-CoT (Ours)	0.765	0.786	0.730	0.695
w/ SC	SylAVL-CoT (Gemini 2.5)	0.760	0.788	0.732	0.697
		Dubbed Lyrics ⟷ Translated			
w/o SC	Google Translate	0.672	0.649	0.618	0.600
w/o SC	mBART-50	0.576	0.634	0.608	0.612
w/o SC	Qwen2.5-72B	**0.677**	**0.655**	0.637	0.633
w/o SC	GPT-4o	0.670	0.635	0.633	0.624
w/o SC	Gemini	0.671	0.652	0.636	0.610
w/ SC	Qwen2.5-72B	0.673	0.632	0.645	0.649
w/ SC	GPT-4o	0.672	0.654	0.647	0.672
w/ SC	Gemini	0.654	0.634	**0.651**	0.669
w/ SC	SylAVL-CoT (Ours)	0.656	0.640	0.647	0.681
w/ SC	SylAVL-CoT (Gemini 2.5)	0.649	0.641	0.645	**0.688**

Table 4: **Comparison with other models: Semantic scores.** The abbreviations in the table are the same as those in Table 3.

similarity using Levenshtein distance (Levenshtein, 1966), as follows:

$$D = \mathrm{LD}(\mathrm{IPA}_{gt}, \mathrm{IPA}_{pred}), \tag{4}$$

where $\mathrm{LD}(\cdot, \cdot)$ refers to the Levenshtein distance function.

4 Experiments

We experiment to validate the effectiveness of `SylAVL-CoT` and analyze the **MAVL** dataset, presenting both quantitative and qualitative results.

4.1 Experimental Setup

Baseline Models. To compare and analyze text-based translation with our model, we selected five models capable of multilingual translation. For traditional multilingual models, we adopted Google Translate (Google) and mBART-50 (Cho et al., 2014), while for LLM models, we selected Qwen2-72B (Yang et al., 2024), GPT-4o (Hurst et al., 2024)[3], and Gemini 2.0 Flash (DeepMind, 2024)[4] with and without syllable constraints. Importantly, all baseline models were evaluated using text-only inputs (original lyrics and syllable constraints when applicable), while our `SylAVL-CoT` approach leverages all three modalities—text, audio, and video—simultaneously through Gemini's multimodal capabilities. Additionally, we evaluated the performance of our `SylAVL-CoT` approach using the more recent Gemini 2.5 Flash model to

[3] gpt-4o-2024-08-06
[4] gemini-2.0-flash-001

assess the impact of model evolution on translation quality.

Evaluation Strategy. We conduct evaluations by comparing the translated lyrics with the original and dubbed languages. The comparison with the original language assesses how faithfully the translation preserves the essence of the source lyrics, while the comparison with the dubbed language evaluates how well the translation reflects linguistic differences in the target language.

4.2 Comparative Analysis

In this section, we evaluate how well the translation models incorporate the three principles described in the Section 3.3. Additional experiments results can be found in Appendix C.

Singability. As shown in Table 3, **`SylAVL-CoT`** achieves the lowest *Syllable Error*, *Syllable Distance*, and *Error Rate* compared to both traditional machine translation models and LLM–based translation. Notably, when using the more recent Gemini 2.5 Flash model, `SylAVL-CoT` demonstrates further improvements across syllable-related metrics. This consistent improvement across languages highlights how our approach can leverage advances in underlying model capabilities.

Sense. Table 4 shows that some machine translation models yield high semantic similarity scores with the original English lyrics, reflecting largely literal translations. In contrast, **`SylAVL-CoT`** shows noticeably lower similarity because it—like human experts—employs extensive paraphrasing and restructuring to maintain singability, thus deviating more from the original wording. This is shown in Figure 4.

When compared to human-translated dubbed lyrics, models with high similarity to the English original drop sharply, showing literal translations miss the creative adjustments in professional dubbing. Meanwhile, **`SylAVL-CoT`** maintains similarity scores to the dubbed lyrics that rival other models, indicating its ability to adopt strategies similar to human translators. Thus, while **`SylAVL-CoT`** sacrifices surface-level closeness to English, it preserves the deeper sense by aligning with how professionals adapt content for singability. More detailed experiments on *Sense* can be found in Appendix C.

Naturalness. Assuming that both the original lyrics and human-translated dubbing uphold a high level of *Naturalness*, we measure how closely each model's output aligns phonetically with these two references. Notably, **`SylAVL-CoT`** yields consistently lower phonetic distance scores across both references, indicating that its generated lyrics share key articulatory and rhythmic traits with professionally produced lyrics. By preserving these phonetic elements, **`SylAVL-CoT`** appears to retain a degree of naturalness similar to the original and dubbed lyrics.

	Model	Levenshtein Distance↓			
		ES	FR	KO	JA
		English Lyrics ⟷ Translated			
	Human Expert	23.22	24.41	25.07	25.40
w/o SC	Google Translate	26.25	26.70	34.29	35.65
w/o SC	MBart-50	50.38	26.83	35.71	32.13
w/o SC	Qwen2.5-72B	26.62	26.89	33.24	31.78
w/o SC	GPT-4o	26.92	27.18	32.29	32.68
w/o SC	Gemini	27.06	27.59	33.90	33.30
w/ SC	Qwen2.5-72B	25.82	25.81	29.22	28.60
w/ SC	GPT-4o	23.51	**23.73**	26.11	26.73
w/ SC	Gemini	24.10	24.97	26.70	26.63
w/ SC	SylAVL-CoT (Ours)	23.16	24.07	**25.20**	**25.88**
w/ SC	SylAVL-CoT (Gemini 2.5)	**23.14**	24.13	25.23	25.90
		Dubbed Lyrics ⟷ Translated			
w/o SC	Google Translate	20.32	22.32	30.79	30.29
w/o SC	MBart-50	44.89	23.09	32.63	26.34
w/o SC	Qwen2.5-72B	20.44	22.30	28.96	25.43
w/o SC	GPT-4o	20.79	22.81	28.50	26.15
w/o SC	Gemini	20.91	22.66	29.07	26.86
w/ SC	Qwen2.5-72B	19.49	21.65	24.65	21.11
w/ SC	GPT-4o	16.62	**19.30**	20.98	18.29
w/ SC	Gemini	17.86	21.22	22.14	18.21
w/ SC	SylAVL-CoT (Ours)	**16.25**	20.07	**19.98**	15.54
w/ SC	SylAVL-CoT (Gemini 2.5)	16.28	20.05	20.04	**15.41**

Table 5: **Comparison with other models: Phonetic distance.** The abbreviations in the table are the same as those in Table 3. **Best** and second are highlighted.

Model	MPNet↑			
	ES	FR	KO	JA
	English Lyrics ⟷ Translated Lyrics			
T	0.7584	0.7888	0.7391	0.6965
T + V	0.7586	0.7794	0.7286	0.6952
T + A	**0.7723**	**0.8020**	**0.7484**	**0.7133**
T + A + V (Ours)	0.7652	0.7859	0.7298	0.6953
	Dubbed Lyrics ⟷ Translated Lyrics			
T	0.6481	0.6332	0.6398	0.6748
T + V	0.6483	0.6284	0.6433	0.6796
T + A	0.6559	0.6400	0.6462	0.6785
T + A + V (Ours)	**0.6561**	**0.6402**	**0.6466**	**0.6814**

Table 6: **Result of ablation study.** "T", "A", and "V" reprsent text, audio and video, repectively.

4.3 Ablation Studies

Table 6 presents an ablation study examining the individual contributions of text, audio, and video inputs in our Multimodal CoT approach. When comparing semantic similarity with the original English lyrics, using only the audio modality yields the highest scores across all target languages.

In contrast, relying solely on video diminishes semantic similarity, suggesting that visual context prompts the model to paraphrase more freely, thus deviating further from the source text. However,

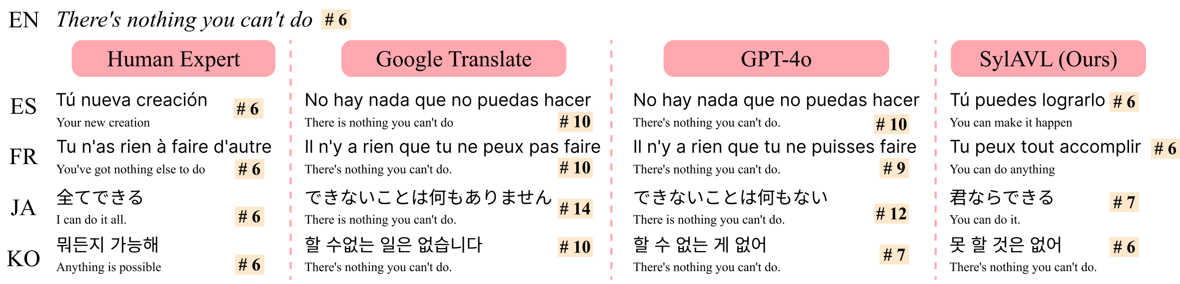

Figure 4: **Qualitative Results.** "#" denotes syllable counts. The English text below each language represents its meaning. This example is "What else can I do?" from "Encanto" by Disney.

Model		Syllable Error↓			
Syllable List	Refine	ES	FR	KO	JA
		English Lyrics ⟷ Translated Lyrics			
✗	✗	3.641	1.860	3.359	6.069
✗	✓	2.805	1.047	1.528	2.955
✓	✗	1.353	0.986	0.879	3.267
✓	✓	**0.966**	**0.902**	**0.695**	**2.572**
		Dubbed Lyrics ⟷ Translated Lyrics			
✗	✗	2.740	2.036	3.244	5.821
✗	✓	2.122	**1.334**	1.711	3.264
✓	✗	1.530	1.337	1.263	3.683
✓	✓	**1.349**	1.363	**1.261**	**3.107**

Table 7: **Result of ablation study.** "✓" and "✗" represent whether the "Syllable List" and "Refining" steps are used or not, respectively.

when we compare outputs against the dubbed lyrics instead of the original English, combining both audio and video yields the best performance. Consequently, leveraging **all modalities** produces the most dubbing-like translations, striking an effective balance between literal accuracy and context-driven paraphrasing.

Table 7 demonstrates that incorporating the "Syllable List" generation and "Refine" stages into our CoT process significantly reduces syllable errors when compared to the original English lyrics. In particular, the introduction of the "Syllable List" stage shows a marked improvement in reducing syllable differences across many languages compared to baselines without it. A similar trend of improvement is observed when comparing the translated lyrics against the dubbed lyrics, with the combination of both stages generally yielding the lowest *Syllable Error*. The prompt used for CoT without "Syllable List" or "Refine" can be found in Appendix F.

4.4 Model Evolution Impact

To assess how advances in pretrained models affect our approach, we compared the performance of **SylAVL-CoT** using different versions of Gemini. As shown in Tables 3 and 4, the transition from Gemini 2.0 Flash to 2.5 Flash yields consistent improvements. The newer model achieves better syllable constraint adherence while maintaining or slightly improving semantic similarity scores. This demonstrates that our **SylAVL-CoT** methodology can effectively leverage improvements in underlying model capabilities without modification, and the MAVL benchmark serves as a valuable tool for evaluating multimodal language models' ability to handle constrained generation tasks.

Furthermore, we investigated whether Gemini performs well with our CoT approach compared to other LLMs. As shown in Table 8, we compared Qwen2.5-72B, GPT-4o, and Gemini with and without CoT prompting using text-only inputs to isolate the effect of CoT reasoning. While GPT-4o achieve similar syllable constraint adherence in their non-CoT configurations, Gemini demonstrates dramatically superior performance when using CoT.

This disparity reveals fundamental differences in how these models process complex, multi-constraint tasks. Qwen2.5 and GPT-4o appear optimized for direct translation but struggle to maintain syllable constraints when reasoning through steps. In contrast, Gemini excels at following structured reasoning paths that involve tracking multiple constraints simultaneously. These findings validate our choice of Gemini for **SylAVL-CoT** and demonstrate that successful multimodal lyrics translation requires not just language understanding, but also the ability to maintain complex constraints throughout a reasoning process.

4.5 User Study

We conducted a user study to evaluate lyric translations generated by various models. All models evaluated, except for Google Translate, are syllable-constrained. More details regarding our user study

Model	CoT	Syllable Error ↓				MPNet Score ↑			
		ES	FR	KO	JA	ES	FR	KO	JA
English Lyrics ⟷ Translated Lyrics									
Qwen2.5-72B	w/o CoT	6.512	2.729	6.502	10.061	**0.883**	**0.878**	**0.823**	**0.810**
	w/ CoT	3.440	1.545	5.785	4.989	0.834	0.844	0.805	0.777
GPT-4o	w/o CoT	3.164	1.442	3.084	7.221	0.830	0.842	0.793	0.782
	w/ CoT	1.888	0.981	1.435	3.538	0.816	0.839	0.785	0.760
Gemini	w/o CoT	3.585	1.384	3.039	6.257	0.781	0.790	0.754	0.733
	w/ CoT	**0.976**	**0.848**	**0.717**	**2.026**	0.758	0.789	0.739	0.697
Dubbed Lyrics ⟷ Translated Lyrics									
Qwen2.5-72B	w/o CoT	5.185	2.893	6.229	9.703	**0.673**	0.632	0.645	0.649
	w/ CoT	2.982	1.792	5.021	5.944	0.655	0.652	0.626	0.654
GPT-4o	w/o CoT	2.403	1.723	3.067	6.940	0.672	**0.654**	**0.647**	**0.672**
	w/ CoT	1.723	**1.219**	1.691	3.820	0.667	0.653	0.675	0.655
Gemini	w/o CoT	2.610	1.703	2.941	5.996	0.654	0.634	0.651	0.669
	w/ CoT	**1.347**	1.297	**1.287**	**2.637**	0.648	0.633	0.640	0.675

Table 8: **Impact of SylAVL-CoT on different LLMs.** The performance of Qwen2.5-72B, GPT-4o, and Gemini with and without **SylAVL-CoT** prompting (text-only, no multimodal inputs).

Language	Model	Singability	Sense	Overall Quality
Spanish	Human Expert	3.90±1.11	3.46±1.20	3.61±1.15
	Google Translate	2.12±1.27	2.60±1.36	2.18±1.25
	Qwen2.5-72B	2.94±1.24	**4.05**±1.00	3.26±1.13
	GPT-4	3.25±1.18	3.90±1.15	3.43±1.13
	Gemini	3.16±1.06	3.22±1.06	3.03±0.98
	SylAVL-CoT	**3.68**±1.16	3.46±1.34	**3.57**±1.26
French	Human Expert	3.94±1.23	3.55±1.17	3.50±1.23
	Google Translate	3.39±1.17	3.80±1.05	3.57±1.32
	Qwen2.5-72B	3.85±1.17	**3.89**±0.95	3.85±1.23
	GPT-4	3.86±1.05	3.86±1.01	3.86±1.14
	Gemini	3.86±1.18	3.65±1.07	3.70±1.19
	SylAVL-CoT	**4.04**±1.03	3.86±0.98	**3.93**±1.11
Korean	Human Expert	3.88±1.24	2.94±1.44	3.28±1.28
	Google Translate	2.05±1.23	2.54±1.39	2.15±1.21
	Qwen2.5-72B	2.00±1.34	3.36±1.12	2.46±1.04
	GPT-4	3.41±1.10	3.33±1.23	3.19±1.07
	Gemini	3.47±1.22	3.29±1.41	3.25±1.26
	SylAVL-CoT	**4.32**±0.81	**3.71**±1.24	**3.95**±1.08
Japanese	Human Expert	3.89±0.80	3.43±0.96	3.57±1.02
	Google Translate	2.21±1.11	2.56±1.13	2.28±1.13
	Qwen2.5-72B	3.17±1.01	3.31±0.81	3.19±0.96
	GPT-4	3.15±0.98	3.39±0.76	3.19±0.98
	Gemini	3.36±1.01	**3.65**±0.77	3.33±1.04
	SylAVL-CoT	**3.84**±0.79	3.60±0.88	**3.64**±0.84

Table 9: **Result of user study.** Mean scores (±standard deviation) from native speakers ($N = 10$ per language) evaluating lyric translations by various models (all syllable-constrained except Google Translate). Metrics include *Singability*, *Sense*, and *Overall Quality*. See Appendix C.1 for details.

methodology, including participant recruitment and task design, can be found in Appendix C.1.

Results in the Table 9 indicate that **SylAVL-CoT** achieved higher *Overall Quality* scores than other models across all languages. Looking at specific aspects, **SylAVL-CoT** demonstrated notably high scores in *Singability* for most languages. However, in French, **SylAVL-CoT**'s advantage in *Singability* over other syllable-constrained models was relatively smaller. This observation aligns with Table 3, which reportedly shows that other models already exhibit lower *Syllable Error* in French compared to other languages, making **SylAVL-CoT**'s lead less pronounced. Furthermore, regarding *Sense*, **SylAVL-CoT** managed to maintain scores comparable to, and sometimes better than, other models, even while prioritizing *Singability*. This balance ultimately contributed to its superior *Overall Quality*.

5 Conclusion

In this paper, we introduced Multilingual Audio-Video Lyrics Benchmark (**MAVL**), the first multilingual and multimodal parallel lyrics translation benchmark that integrates text, audio, and video for singable translations. We also proposed **SylAVL-CoT**, which leverages existing MLLMs without fine-tuning and enforces syllable constraints through Chain-of-Thought reasoning. Our experiments show that **SylAVL-CoT** balances expressive paraphrasing with contextual accuracy, addressing a key gap in musical animations. We hope these advances pave the way for new automated lyrics translation systems and further research in multilingual, multimodal machine translation.

6 Limitations

While the proposed **MAVL** dataset and the **SylAVL-CoT** framework address several challenges in multilingual, multimodal lyrics translation, there are still limitations:

Data Scope. Our dataset mainly focuses on animated musicals and on five target languages (Spanish, French, Japanese, and Korean, in addition to the original English). Although the dataset provides a rich testing ground for multimodal translation, their thematic variety may not represent the full range of musical genres, languages, and styles encountered in broader contexts. Consequently, models tested solely on this data may not generalize well to other genre-specific lyrical structures, or under-represented low-resource languages.

However, we view MAVL as a foundational step toward broader coverage. The dataset collection pipeline and quality standards established here can serve as a blueprint for expanding to other genres. For instance, the insights gained from MAVL could be leveraged to develop automated quality assessment models that identify high-quality translations in genres like pop music or K-pop, where fan-made translations are abundant but vary in quality. Additionally, our framework could be extended to semi-automatically curate datasets for low-resource languages by combining professional translations where available with carefully filtered community contributions.

Line-based translation. Our current `SylAVL-CoT` approach primarily translates lyrics on a line-by-line basis. However, effective lyric translation often benefits from more flexible strategies, such as rephrasing across line breaks, or splitting and merging lines, which can significantly enhance singability and poetic expression. Such holistic strategies are often best implemented at a section level rather than a strict line level, allowing for more natural rhythmic and semantic flow. While the **MAVL** dataset includes section-level annotations, our current model does not fully leverage this. We anticipate that future research could utilize these section-level annotations to explore more sophisticated, context-aware translation strategies that transcend single-line processing, leading to more natural and musically-fitting translations.

Tonal Language Application. Our framework, aiming for broad multilingual applicability, does not currently incorporate specialized mechanisms for tonal languages such as Chinese. As highlighted in (Guo et al., 2022) and (Ye et al., 2024), translating lyrics into tonal languages often requires specific considerations for tone contours to preserve musicality and meaning, which can involve distinct processing steps. Our pursuit of a general-purpose solution meant these language-specific tonal constraints were not a primary focus. We hope that future iterations of our work can be extended to address the unique challenges of tonal languages, potentially by integrating or adapting techniques from existing research to enhance performance in these linguistic contexts.

Alignment Challenges. Precise synchronization of lyrics, audio, and video in musical settings remains non-trivial. Although we employ techniques such as Whisper-based alignment and careful human annotation, discrepancies can persist, especially for lines containing overlapping voices, spoken dialogue, or onomatopoeic interjections. These alignment inaccuracies may lead to sub-optimal multimodal model training or evaluation. Future work could incorporate more robust audio-visual alignment methods or user-in-the-loop correction to refine time stamping for each lyric segment.

Evaluation Metrics. Although our evaluation framework focuses on singability, sense, naturalness, these metrics still cannot completely capture musicality or artistic style. Automated metrics do not fully reflect subjective audience judgments. Additionally, cultural references and emotional nuance might be lost in translation and remain difficult to quantify objectively. Future work could explore LLM-based evaluation frameworks to better capture these subjective aspects, potentially using multi-agent systems where different agents assess rhythmic fit, emotional alignment, and poetic quality.

Broader Applicability. Our emphasis on animated musical translations may not translate directly to other domains such as live theater, opera, pop music, or user-generated musical content. The complexities in live performances, spontaneous improvisations, or multi-speaker settings are beyond the current benchmark's scope. Future research could extend the approach to a wider range of musical and performance contexts to validate the model's robustness and adaptability.

Overall, these limitations highlight the need for broader, more diversified datasets, refined alignment techniques, and more holistic metrics to capture the creative and performative aspects of lyric translation. We hope that releasing **MAVL** and proposing `SylAVL-CoT` spark further innovations and encourage the research community to build on or address these limitations in future work.

7 Ethical Consideration

In conducting this research, we have taken various ethical aspects into account to ensure responsible and fair practices in the development and dissemination of our work.

Transparency. We are committed to maintaining transparency throughout our research process. All preprocessing steps, alignment techniques, and model training methodologies are fully disclosed to enable reproducibility and facilitate further scrutiny by the research community. More details are in Appendix.

Copyright Compliance. To respect intellectual property rights, our dataset does not distribute copyrighted materials directly. Instead, we provide structured metadata and download links where applicable, ensuring compliance with copyright regulations while preserving the dataset's usability for research.

Cultural Sensitivity and Inclusivity. Our research aims to contribute to a diverse and inclusive representation of musical content across languages. We have taken measures to respect cultural nuances and avoid biases, ensuring that our approach promotes fairness in singable lyric translation.

Potential Societal Impact. We acknowledge that our work may have broader implications for cross-cultural communication and creative industries. We encourage continued ethical reflection on how AI-assisted lyric translation can be leveraged responsibly, particularly in artistic and commercial applications.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korean government (MSIT) (No. RS-2024-00457882, AI Research Hub Project) and by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (Nos. RS-2024-00354218 and RS-2024-00353125).

References

Rocío Baños Piñero and Frederic Chaume. 2009. Prefabricated orality: A challenge in audiovisual translation.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Third Conference on Machine Translation (WMT18)*, volume 2, pages 308–327.

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

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *Preprint*, arXiv:1409.1259.

Google DeepMind. 2024. Gemini 2.0 flash. Accessed: 2025-02-15.

Yexing Du, Ziyang Ma, Yifan Yang, Keqi Deng, Xie Chen, Bo Yang, Yang Xiang, Ming Liu, and Bing Qin. 2024. Cot-st: Enhancing llm-based speech translation with multimodal chain-of-thought. *Preprint*, arXiv:2409.19510.

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

Google. Google translate. Accessed: 2025-02-15.

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.

Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. 2024. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18180–18187.

Ke Hu, Zhehuai Chen, Chao-Han Huck Yang, Piotr Żelasko, Oleksii Hrinchuk, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024a. Chain-of-thought prompting for speech translation. *arXiv preprint arXiv:2409.11538*.

Tianxiang Hu, Pei Zhang, Baosong Yang, Jun Xie, Derek F. Wong, and Rui Wang. 2024b. Large language model for multi-domain translation: Benchmarking and domain cot fine-tuning. *Preprint*, arXiv:2410.02631.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

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

John Laver. 1994. Principles of phonetics. *Cambridge: Cambridge Univ*.

Chrisna Leni and Athriyana Santye Pattiwael. 2019. Analyzing translation strategies utilized in the translation of song "do you want to build a snowman?". *Journal of Language and Literature*.

VI Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics doklady*.

Chengxi Li, Kai Fan, Jiajun Bu, Boxing Chen, Zhongqiang Huang, and Zhi Yu. 2023. Translate the beauty in songs: Jointly learning to align melody and translate lyrics. *Preprint*, arXiv:2303.15705.

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

Peter Low. 2005. The pentathlon approach to translating songs. In *Song and significance*, pages 185–212. Brill.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *Preprint*, arXiv:2501.07246.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Longshen Ou, Xichu Ma, Min-Yen Kan, and Ye Wang. 2023. Songs across borders: Singable and controllable neural lyric translation. *Preprint*, arXiv:2305.16816.

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.

Olena Pidhrushna. 2021. Functional approach to songs in film translation: Challenges and compromises.

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orašan, Tharindu Ranasinghe, and Fred Blain. 2024. What do large language models need for machine translation evaluation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Pawan Rajpoot, Nagaraj Bhat, and Ashish Shrivastava. 2024. Multimodal machine translation for low-resource indic languages: A chain-of-thought approach using large language models. In *Proceedings of the Ninth Conference on Machine Translation*, pages 833–838.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Preprint*, arXiv:2004.09297.

Moh Supardi and Dea Putri. 2018. Audio-visual translation techniques: Subtitling and dubbing of movie soundtrack in frozen: Let it go. *Buletin Al-Turas*, 24:399–414.

Şebnem Susam-Sarajeva. 2008. Translation and music. *The Translator*, 14:187–200.

Christopher Taylor. 2016. The multimodal approach in audiovisual translation. *Target*, 28.

Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. 2024. Videocot: A video chain-of-thought dataset with active annotation tool. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 92–101.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-cot: Let vision language models reason step-by-step. *ArXiv*, abs/2411.10440.

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

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.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Preprint*, arXiv:2302.00923.

A Explanations of the choice for the Metrics

Our evaluation approach enhances traditional methods by incorporating comparisons not only with original lyrics but also with actual dubbed translated lyrics. This dual comparison better reflects cultural and linguistic nuances and critically assesses whether the translated lyrics can be articulated naturally in the target language. The rationale for selecting each evaluation metric is detailed below.

A.1 Evaluation for "Singability" (Syllable-focused)

We prioritized syllable-based measurements for singability as they more directly reflect the difficulty of singing compared to phoneme-level analyses. This approach is supported by previous studies (Guo et al., 2022; Ye et al., 2024; Kim et al., 2024), which have demonstrated the effectiveness of syllable-level analysis in assessing singing performance.

Syllable Error (SE) As noted by (Ye et al., 2024), an increase in the number of syllables to be pronounced generally makes singing more challenging. To capture this, we employed a weighted method using the parameter β. The choice of β is crucial for accurately reflecting singing difficulty. Following (Ye et al., 2024), we set $\beta = 2$. This specific value models the principle that singing difficulty increases more than linearly (proportionally, in this context implying a significant impact) with an increasing number of syllables. A higher β penalizes excessive syllables more heavily, which aligns with the practical observation that cramming too many syllables into a musical phrase significantly degrades singability. Thus, $\beta = 2$ is a critical setting for evaluating how well the translated lyrics maintain a singable syllable count.

Syllable Count Distance (SCD) Proposed by (Kim et al., 2024), SCD measures the congruity between the translated lyrics and the original structure. It achieves this by calculating a normalized relative distance based on the absolute differences in syllable counts, considering both the original-to-translation and translation-to-original directions. This metric is valuable for assessing not only translation accuracy in terms of length but also the consistency of rhythmic structure between the source and target lyrics.

Error Rate This metric provides a straightforward measure of singability by quantifying the proportion of incorrect syllables relative to the correct or reference values. Its simplicity offers an intuitive way to evaluate overall singability and helps in identifying common error patterns in syllable mapping.

A.2 Evaluation for "Sense"

We opted for deep learning-based methods to evaluate "Sense" because traditional metrics like BLEU and METEOR, which rely on n-gram overlap or word matching, are often inadequate for lyric translation. Lyrics frequently require creative adaptation rather than literal translation to preserve the song's intended meaning, emotional impact, and artistic essence. Deep learning models are better equipped to assess these nuanced translations by considering contextual meaning.

MPNet-based semantic score To evaluate semantic accuracy across multiple languages, we utilized state-of-the-art deep learning models. Specifically, we employed multilingual sentence transformers from the SBERT library, which represent the current leading technology for assessing multilingual semantic similarity. This allows for a more robust evaluation of whether the core meaning of the lyrics is preserved post-translation. We also explored alternative metrics, the results of which are detailed in Table 11.

A.3 Evaluation for "Naturalness"

The naturalness of translated lyrics is paramount for their acceptance and performance. We assess this through phonetic similarity.

Phonetic Distance To quantify "Naturalness," we measure the phonetic similarity using Levenshtein distance calculated on the International Phonetic Alphabet (IPA) transcriptions of the translated lyrics and the reference lyrics (either original or dubbed). This serves as an effective proxy for naturalness for two main reasons:

1. It quantifies the phonetic deviation from reference lyrics, which are assumed to be inherently natural and pronounceable in their respective languages. A lower distance suggests that the translation inherits this natural phonetic structure, making it more likely to sound fluent.

2. Higher phonetic similarity (i.e., lower Levenshtein distance) implies that the translated lyrics are easier to pronounce and possess a smoother phonetic flow, mirroring the articulatory ease of the reference lyrics. This contributes significantly to the perceived naturalness of the translation when sung.

B Analysis on the Dataset

B.1 Similarity Distribution across Languages

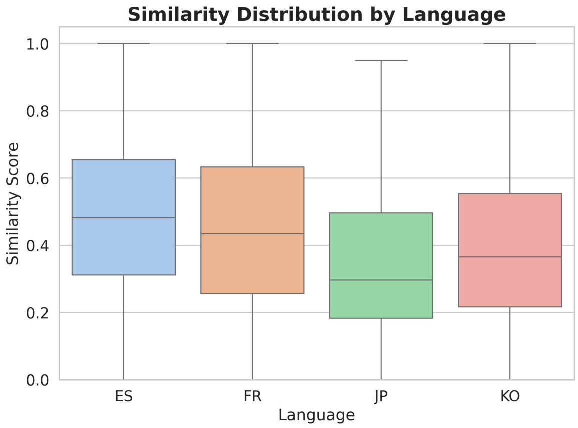

Figure 5: **MPNet Similarity Distribution By Languages.**

Figure 5 illustrates the MPNet similarity distribution between the original English lyrics and their dubbed counterparts across various languages. It is important to note the methodology used for this specific visualization. Unlike the multilingual MPNet approach potentially discussed elsewhere, here we aimed to mitigate biases arising from inherent linguistic distances. Such distances could skew similarity scores even when translations are relatively literal, potentially misrepresenting the true degree of translational fidelity. Therefore, the dubbed lyrics for each language (ES, FR, JP, KR) were first translated into English using Google Translate. Subsequently, the similarity between these English-translated dubbed lyrics and the original English lyrics was computed using an English-specific MPNet[5]. This approach was chosen to enhance the reliability of the analysis; however, it is worth noting that when directly measuring the similarity between the original English lyrics and the dubbed lyrics using a multilingual MPNet, the results were consistent with the findings presented here. The chart reveals distinct similarity patterns, with scores generally following the order: ES, FR, KO, and then JA. This observation suggests that languages like Spanish (ES) and French (FR), which share more structural and alphabetical similarities with English, tend to feature more literal translations, resulting in higher similarity scores. Conversely, East Asian languages, particularly Japanese (JA), exhibit lower similarity. This indicates that their dubbed versions likely contain more significantly different expressions or a greater degree of free translation, reflecting the substantial linguistic divergence from English. These findings are also corroborated by the Human Expert Row in Table 4.

[5]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

C Additional Experiments Details

C.1 User Study Details

For user study, we developed a dedicated website where participants could watch the 10 original English video clips alongside the translated lyrics from different systems. To ensure a clear understanding of the evaluation task and to guide participants towards evaluations aligned with our intended criteria, we provided comprehensive instructions before they began. These instructions included a detailed outline of the tasks to be performed, clear definitions and criteria for each evaluation metric, and examples. This preparation aimed to equip participants to make informed and consistent judgments.

Specifically, before commencing the evaluation, participants were presented with an "Evaluation Guide." This guide, titled "Evaluation Guide," first outlined its purpose: "On this page, you will evaluate the quality of song lyrics translation." It then provided the following key instructions:

- Compare the original lyrics with the translated lyrics.
- Evaluate the translated lyrics along with the melody of the song through the provided video.

Participants were asked to evaluate the translations based on three criteria:

- **Singability:** Evaluate how well the translated lyrics fit with the melody. If possible, try singing them yourself.
- **Sense:** Evaluate how clearly and accurately the translated lyrics convey the original meaning.

	Model	MiniLM				MPNet				BERTScore			
		ES	FR	KO	JA	ES	FR	KO	JA	ES	FR	KO	JA
		English Lyrics ⟷ Translated Lyrics											
	Human Expert	0.1804	0.1622	0.0755	0.0801	0.6392	0.6126	0.5746	0.5374	0.6856	0.6867	0.6486	0.6383
w/o SC	Google Translate	**0.2290**	0.2322	**0.0912**	0.0810	**0.9052**	**0.9027**	**0.8572**	**0.8569**	**0.7751**	**0.7811**	0.6805	0.6743
w/o SC	MBart-50	0.1993	**0.2491**	0.0856	**0.0994**	0.7658	0.8895	0.8496	0.8415	0.7331	0.7775	0.6677	**0.6828**
w/o SC	Qwen-72B	0.2174	0.2258	0.0657	0.0726	0.9004	0.8987	0.8412	0.8362	0.7725	0.7771	**0.6849**	0.6784
w/o SC	GPT-4o	0.2131	0.2167	0.0628	0.0706	0.8993	0.8969	0.8486	0.8422	0.7688	0.7727	0.6812	0.6747
w/o SC	Gemini	0.2061	0.2146	0.0687	0.0725	0.8931	0.8911	0.8426	0.8337	0.7575	0.7624	0.6747	0.6621
w/ SC	Qwen-72B	0.2042	0.2046	0.0653	0.0778	0.8825	0.8776	0.8226	0.8103	0.7615	0.7656	0.6812	0.6764
w/ SC	GPT-4o	0.2042	0.2047	0.0703	0.0837	0.8295	0.8417	0.7932	0.7818	0.7421	0.7542	0.6759	0.6711
w/ SC	Gemini	0.1857	0.1785	0.0673	0.0826	0.7813	0.7904	0.7544	0.7327	0.7169	0.7234	0.6686	0.6604
w/ SC	SylAVL-CoT (Ours)	0.1900	0.1849	0.0679	0.0867	0.7652	0.7859	0.7289	0.6953	0.7142	0.7278	0.6635	0.6567
		Dubbed Lyrics ⟷ Translated Lyrics											
w/o SC	Google Translate	0.5584	0.5092	0.6661	0.4619	0.6724	0.6490	0.6177	0.5995	0.7402	0.7348	0.7158	0.6824
w/o SC	MBart-50	0.4938	0.4891	0.6671	0.4547	0.5764	0.6339	0.6075	0.6123	0.7031	0.7258	0.7036	0.6925
w/o SC	Qwen-72B	0.5622	0.5145	0.6920	0.4820	**0.6772**	**0.6549**	0.6371	0.6329	0.7423	**0.7384**	0.7359	0.7012
w/o SC	GPT-4o	0.5564	0.5001	0.6932	0.4872	0.6703	0.6354	0.6332	0.6244	0.7386	0.7321	0.7353	0.7005
w/o SC	Gemini	**0.5623**	**0.5200**	0.6946	0.4794	0.6709	0.6516	0.6359	0.6104	0.7328	0.7332	0.7318	0.6885
w/ SC	Qwen-72B	0.5619	0.4934	0.6993	0.4905	0.6729	0.6323	0.6448	0.6489	0.7410	0.7309	0.7416	0.7118
w/ SC	GPT-4o	0.5573	0.5107	**0.7085**	**0.5042**	0.6722	0.6536	0.6472	0.6715	**0.7450**	0.7380	0.7530	0.7238
w/ SC	Gemini	0.5383	0.4910	0.6979	0.4871	0.6538	0.6341	**0.6506**	0.6687	0.7314	0.7224	0.7468	0.7152
w/ SC	SylAVL-CoT (Ours)	0.5395	0.4975	0.7016	0.4975	0.6561	0.6402	0.6467	**0.6814**	0.7358	0.7300	**0.7509**	**0.7241**

Table 10: **Comparison with other models: Sense. Best** are highlighted.

Modality	MiniLM				MPNet				BERTScore			
	ES	FR	KO	JA	ES	FR	KO	JA	ES	FR	KO	JA
	English Lyrics ⟷ Translated Lyrics											
T	0.1872	0.1846	**0.0692**	0.0894	0.7584	0.7888	0.7391	0.6965	0.7082	0.7265	0.6651	0.6592
T + V	0.1871	0.1788	0.0692	0.0849	0.7586	0.7794	0.7286	0.6952	0.7098	0.7236	0.6630	0.6572
T + A	**0.1921**	**0.1874**	0.0688	**0.0897**	**0.7723**	**0.8020**	**0.7484**	**0.7133**	**0.7143**	**0.7320**	**0.6661**	**0.6608**
T + A + V (Ours)	0.1900	0.1849	0.0679	0.0867	0.7652	0.7859	0.7298	0.6953	0.7142	0.7278	0.6635	0.6567
	Dubbed Lyrics ⟷ Translated Lyrics											
T	0.5252	0.4869	0.6969	0.4874	0.6481	0.6332	0.6398	0.6748	0.7296	0.7245	0.7480	0.7243
T + V	0.5346	0.4878	0.6995	**0.4929**	0.6483	0.6284	0.6433	0.6796	0.7315	0.7231	0.7489	0.7256
T + A	0.5382	0.4918	0.6995	0.4895	0.6559	0.6400	0.6462	0.6785	0.7345	0.7278	0.7500	**0.7262**
T + A + V (Ours)	**0.5395**	**0.4975**	**0.7016**	0.4893	**0.6561**	**0.6402**	**0.6466**	**0.6814**	**0.7349**	**0.7285**	**0.7509**	0.7260

Table 11: **Combined metrics of Ablation studies for MiniLM, MPNet, and BERTScore.**

EN *Something sharp, something new* # 6

	Human Expert	Google Translate	GPT-4o	SylAVL (Ours)
ES	Forma audaz, ¡Qué impresión! Bold form, what an impression! # 9	Algo agudo, algo nuevo Something sharp, something new # 9	Algo afilado, algo nuevo # 10 Something sharp, something new	Filo agudo hoy # 6 Sharp edge today
FR	C'est piquant, c'est nouveau # 6 It's spicy, it's new	Quelque chose de net, quelque chose de nouveau # 11 There's nothing you can't do.	Quelque chose de tranchant, quelque chose de nouveau # 12 Something sharp, something new	Un truc pointu, du neuf # 6 Something sharp, something new
JA	棘もあるわ # 6 There are thorns, too.	鋭いもの、新しいもの # 12 Something sharp, something new.	何か鋭いもの、新しいもの # 15 Something sharp, something new.	斬新な何か # 7 Something novel
KO	날카롭고 새로워 # 7 Sharp and fresh	날카 롭고 새로운 것 # 8 Something sharp and new	뭔가 날카로운 것, 뭔가 새로운 것 # 13 Something sharp, something new	새로운 짜릿함 # 6 A new sense of excitement

EN *You gotta dig a little deeper* # 9

	Human Expert	Google Translate	GPT-4o	SylAVL (Ours)
ES	Es tu deber, trabaja duro It's your duty, work hard # 9	Tienes que cavar un poco más profundo You have to dig a little deeper # 12	Tienes que excavar un poco más profundo You have to dig a little deeper # 12	Debes buscar más en tu alma You must search more in your sou # 9
FR	Il faut creuser encore et encore # 9 Digging and more digging	Tu dois creuser un peu plus profondément You need to dig a little deeper # 11	Tu dois creuser un peu plus profond You need to dig a little deeper # 9	Il faut aller creuser plus profond Dig deeper # 9
JA	もう一度考えて Think again. # 7	もう少し深く掘り下げなければなり ません # 19 We need to dig a little deeper.	もっと深く掘らなきゃ # 9 We need to dig deeper.	探して、もっと奥を Look for it. Look deeper. # 9
KO	조금만 더 노력을 해 봐 Try a little harder # 9	당신은 조금 더 깊이 파헤쳐 야합니다 You need to dig a little deeper # 15	조금 더 깊이 파야 해 We need to dig a little deeper # 8	좀 더 깊이 파 봐 계속해 Dig a little deeper. Continue # 9

EN *With new horizons to pursue* # 8

	Human Expert	Google Translate	GPT-4o	SylAVL (Ours)
ES	Un horizonte nuevo abrir. A new horizon to open. # 9	Con nuevos horizontes para perseguir With new horizons to pursue # 12	Con nuevos horizontes que perseguir With new horizons to pursue # 11	Tras nuevos mundos sin dudar After new worlds without hesitation # 8
FR	Vers les horizons du bonheur # 8 Towards the horizons of happiness	Avec de nouveaux horizons à poursuivre # 12 With new horizons to pursue	Avec de nouveaux horizons à poursuivre With new horizons to pursue # 11	Vers l'avenir à découvrir Into the future to discover # 8
JA	新しい世界 A New World # 8	追求する新しい地平線で # 12 With new horizons to pursue	新しい地平線を追い求めて Pursuing New Horizons # 13	新たな夢見て # 8 Dream a new dream.
KO	밤하늘 가슴에 안고 # 8 Holding the night sky to your chest	추구 할 새로운 지평이 있습니다 # 13 There are new horizons to pursue	새로운 지평을 추구하며 # 10 Seeking new horizons	새 지평선을 따라서 # 8 There's nothing you can't do.

Figure 6: **Qualitative results**. This figure showcases translations of English lyrics into Spanish (ES), French (FR), Japanese (JA), and Korean (KO) by Human Experts, Google Translate, GPT-4o, and **SylAVL-CoT** (Ours). Notably, the translations from **SylAVL-CoT** (Ours) demonstrate syllable counts (where # denotes the syllable counts) that are most similar to those of the original English lyrics. The examples also allow for a comparison of how specific English lyric lines are rendered by Human Experts versus our **SylAVL-CoT** model.

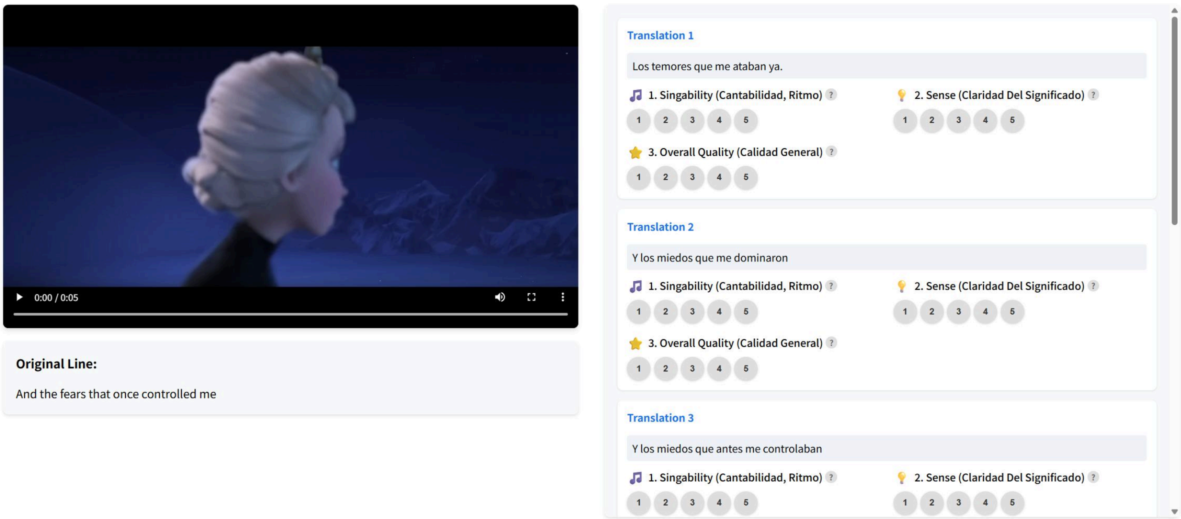

Figure 7: **Example of User Study Website**

- **Overall Quality:** Evaluate the overall quality of the translation.

The scoring mechanism was explained as: "Rate each translation from 1 to 5. (1: Very poor, 2: Poor, 3: Average, 4: Good, 5: Very good)". This entire guide was translated and presented to participants in their respective evaluation languages. Finally, participants would click a "Start Evaluation" button to proceed. Example image of user study website can be seen in Figure 7.

Each evaluation session lasted approximately 30 minutes and involved 10 participants per language. Participants were compensated at a rate of approximately €5 each for their time and effort.

C.2 Qualitative Results

We present various qualitative results, as shown in Figure 6. `SylAVL-CoT` preserves both the original meaning and syllable structure, ensuring singability.

C.3 Semantic Evaluation

Comparison with other models. We conducted comparative experiments using the metrics employed in previous studies to measure semantic fidelity. This is presented in Table 10. For the translation between English and non-English, Google Translate scored the highest score with the overall languages. As human expert considers various factors for translating the lyrics, the performance is the lowest among various semantic evaluation metrics. When comparing dubbed and translated lyrics, our proposed model achieved high performance for a wide range of languages. On the other hand, Google Translate's translation performance seems to be relatively poor.

Ablation study. To demonstrate the usefulness of multimodality, we conducted a semantic evaluation based on different modalities, as shown in Table 11. As observed in Table 11, incorporating multimodal information alongside text leads to better performance.

C.4 Cross-lingual `SylAVL-CoT` Experiments

The MAVL dataset is inherently well-suited for comprehensive cross-lingual benchmarking. Most of its data entries across all supported languages (English, Spanish, French, Korean, and Japanese) include aligned video and meticulously synchronized lyrics. This rich, multimodal, and parallel structure enables the evaluation of translation not only from a common source language (like English) to various target languages but also between any pair of the supported languages, or from non-English languages back to English. This flexibility allows for a deeper understanding of a model's translation capabilities across diverse linguistic landscapes.

To illustrate this, we conducted additional cross-lingual experiments with `SylAVL-CoT`, evaluating its performance when translating between different language pairs beyond the English-centric evaluations presented in the main paper. The Syllable Error (SE), Syllable Distance and MPNet semantic similarity scores for these experiments are presented in Table 12 and Table 14, respectively.

Upon analyzing the results, it is observed that

Metric	Syllable Error ↓															
Source	ES				FR				KO				JA			
Target	EN	FR	KO	JA	EN	ES	KO	JA	EN	ES	FR	KO	EN	ES	FR	JA
English Lyrics ⟷ Translated Lyrics																
Human Expert	1.005	1.280	1.221	1.740	0.824	1.945	1.225	1.772	0.659	1.743	1.099	1.475	1.321	2.239	1.617	1.454
Google Translate	2.154	1.994	7.021	13.106	1.791	5.042	9.306	16.869	2.631	3.114	2.530	8.991	3.155	2.821	2.901	3.833
Gemini	1.442	1.404	1.169	3.443	0.773	3.105	2.021	5.229	1.301	1.634	1.274	4.023	2.152	1.191	1.983	1.280
SylAVL-CoT (Ours)	**0.298**	**0.965**	**0.382**	**1.640**	**0.353**	**1.067**	**0.613**	**2.050**	**0.311**	**0.795**	**0.932**	**1.704**	**0.333**	**0.646**	**1.053**	**0.348**
Dubbed Lyrics ⟷ Translated Lyrics																
Google Translate	1.838	2.265	7.886	14.200	1.770	4.267	9.157	16.662	2.554	3.246	2.732	9.213	2.996	3.096	3.001	4.017
Gemini (w/ SC)	**1.030**	**1.412**	**1.854**	4.718	**0.997**	2.519	2.280	5.339	1.397	1.896	1.534	4.364	2.107	2.033	2.137	1.631
SylAVL-CoT (Ours)	1.780	1.746	1.966	**3.173**	1.062	**1.658**	**1.444**	**2.848**	**1.035**	**1.504**	**1.501**	**2.405**	**1.527**	**1.810**	**1.905**	**1.514**

Table 12: **Comparison of Syllable Error (SE) in Cross-Lingual Lyrics Translation for various Source Languages (ES, FR, KO, JA).**

Metric	Syllable Distance ↓															
Source	ES				FR				KO				JA			
Target	EN	FR	KO	JA	EN	ES	KO	JA	EN	ES	FR	KO	EN	ES	FR	JA
English Lyrics ⟷ Translated Lyrics																
Human Expert	0.116	0.136	0.132	0.172	0.078	0.136	0.153	0.108	0.082	0.132	0.108	0.138	0.133	0.172	0.153	0.138
Google Translate	0.303	0.237	0.609	0.398	0.204	0.300	0.779	0.497	0.474	0.309	0.387	0.484	0.557	0.325	0.438	0.336
Gemini	0.174	0.168	0.202	0.102	0.077	0.182	0.300	0.133	0.156	0.116	0.150	0.235	0.295	0.135	0.277	0.148
SylAVL-CoT (Ours)	**0.028**	**0.104**	**0.109**	**0.029**	**0.027**	**0.072**	**0.138**	**0.045**	**0.030**	**0.058**	**0.105**	**0.129**	**0.041**	**0.053**	**0.130**	**0.040**
Dubbed Lyrics ⟷ Translated Lyrics																
Google Translate	0.251	0.236	0.699	0.461	0.201	0.267	0.773	0.492	0.455	0.352	0.402	0.518	0.531	0.376	0.445	0.361
Gemini (w/ SC)	0.126	0.156	0.290	0.154	0.095	0.165	0.319	0.166	0.167	0.160	0.175	0.271	0.296	0.231	0.296	0.191
SylAVL-CoT (Ours)	**0.120**	**0.153**	**0.221**	**0.144**	**0.086**	**0.146**	**0.216**	**0.128**	**0.146**	**0.192**	**0.165**	**0.266**	**0.138**	**0.177**	**0.207**	**0.150**

Table 13: **Comparison of Syllable Distance (SD) in Cross-Lingual Lyrics Translation for various Source Languages (ES, FR, KO, JA).**

Metric	MPNet Score ↑															
Source	ES				FR				KO				JA			
Target	EN	FR	KO	JA	EN	ES	KO	JA	EN	ES	FR	KO	EN	ES	FR	JA
English Lyrics ⟷ Translated Lyrics																
Human Expert	0.639	0.578	0.580	0.574	0.609	0.578	0.561	0.565	0.582	0.580	0.561	0.617	0.535	0.574	0.565	0.617
Google Translate	**0.884**	**0.909**	**0.857**	**0.863**	**0.897**	**0.920**	**0.871**	**0.879**	**0.812**	**0.837**	**0.839**	**0.865**	**0.781**	**0.824**	**0.831**	0.848
Gemini (w/ SC)	0.809	0.833	0.800	0.798	0.823	0.850	0.803	0.803	0.767	0.799	0.790	0.862	0.736	0.792	0.785	**0.857**
SylAVL-CoT (Ours)	0.762	0.804	0.783	0.757	0.776	0.810	0.779	0.759	0.713	0.768	0.766	0.804	0.691	0.759	0.761	0.837
Dubbed Lyrics ⟷ Translated Lyrics																
Google Translate	0.590	0.561	0.556	0.565	0.563	0.552	0.536	0.550	0.522	0.527	0.517	0.590	0.460	0.498	0.494	0.536
Gemini (w/ SC)	0.597	0.570	**0.600**	0.641	**0.563**	0.551	0.568	0.620	**0.545**	**0.546**	0.530	0.670	0.464	0.504	0.494	0.574
SylAVL-CoT (Ours)	**0.599**	**0.576**	0.598	**0.662**	0.560	**0.552**	**0.574**	**0.644**	0.541	0.544	**0.541**	**0.676**	**0.473**	**0.505**	**0.508**	**0.567**

Table 14: **Comparison of MPNet Score in Cross-Lingual Lyrics Translation for various Source Languages (ES, FR, KO, JA).**

SylAVL-CoT generally demonstrates a performance profile consistent with that reported for English-to-other-language translations in the main body of this paper (refer to Table 3 for syllable error metrics and Table 4 for semantic scores). While minor variations naturally occur depending on the specific linguistic characteristics and distance between language pairs, the overall efficacy of `SylAVL-CoT` in maintaining both singability and semantic coherence remains evident across these broader cross-lingual scenarios. This underscores the robustness of the `SylAVL-CoT` approach and the utility of the MAVL dataset for multifaceted translation evaluation. Notably, when Spanish is the source language, a relatively higher Syllable Error (SE) with Dubbed Lyrics can be observed. This aligns with the observation from Table 3, where Human Expert translations into Spanish also showed a high SE. This suggests that Spanish lyrics tend to have a higher syllable count per line compared to the original English lyrics. Consequently, translations generated by `SylAVL-CoT` from a Spanish source, which aim for low Syllable Error and Syllable Distance against the Spanish source, might naturally reflect this higher syllable count distribution. If the Dubbed Lyrics (against which the comparison is made) have a syllable count closer to the original English (i.e., generally lower), this would explain the increased SE when comparing `SylAVL-CoT`'s Spanish-sourced output to these Dubbed Lyrics.

D Details of MAVL Dataset

D.1 Details of Lyrics Collection by Web Crawling

This section describes how we collected multilingual lyrics and corresponding videos, as illustrated in Figure 8. We began by gathering metadata for animated songs from `last.fm` (Figure 8-(a)), followed by collecting English lyrics from `genius` (Figure 8-(b)). Based on this information, we searched `lyricstranslate.com` by country to collect localized lyrics and corresponding videos in multiple languages (Figure 8-(c)).

D.2 Alignment Methodology

To maximize alignment accuracy between lyrics, audio, and video, we employed a comprehensive approach using `stable-ts`, which builds upon the Whisper model (Radford et al., 2022). We chose stable-ts over alternatives like WhisperX because it is specifically designed for aligning existing text with audio, rather than just creating timestamps, and it provides robust multilingual support critical for our diverse language set.

Our alignment process incorporates several quality assurance techniques:

- **Ensemble Approach**: We used an ensemble of Whisper-large-v1 and v2 models (v3 was found to be less reliable for this task), selecting the best alignment based on confidence scores.
- **Vocal Separation**: We employed the DEMUCS model to separate vocals from background music, performing alignment on both the original track and the vocal-only track. If the original track alignment failed, we used the vocal-only track alignment as a fallback.
- **Quality Filtering**: Alignments with low confidence scores or significant timing discrepancies were excluded from the dataset to ensure high quality.

D.3 MAVL Dataset Format

In order to comply with copyright regulations, we only provide URLs rather than distributing the full lyrics, and we reconstruct each line for alignment by extracting a compact representation. For instance, for the English line “Remember me though I have to say goodbye,” we split it by spaces and record the first letter of each word (R, m, t, I, h, t, s, g) along with the line’s first and last words (“Remember” and “goodbye”), resulting in a representation like [“RmtIhtsg”, “Remember”, “goodbye”]. With Japanese lyrics—where spacing does not naturally separate words—we use morphological analysis (e.g., MeCab[6]) to split the line into tokens. We then combine these tokens in pairs, which we treat similarly to the English case by storing partial data (such as the concatenation of first letters or selected tokens) to enable accurate restoration of the original line once the corresponding URL is accessed.

Upon finalizing line reconstruction, IPA transcription, and syllable counting, the dataset assumes the structure depicted in Figure 9.

E Why we choose CoT method

Building an end-to-end multimodal lyrics translation system that handles text, audio, and video poses several unique challenges. Multilingual lyrics datasets with aligned audiovisual content are extremely rare. Training a dedicated model from scratch on such limited data is a significant hurdle. Even if such a model were developed, the training process itself would be time-consuming and resource-intensive. Furthermore, adapting the

[6]https://github.com/SamuraiT/mecab-python3

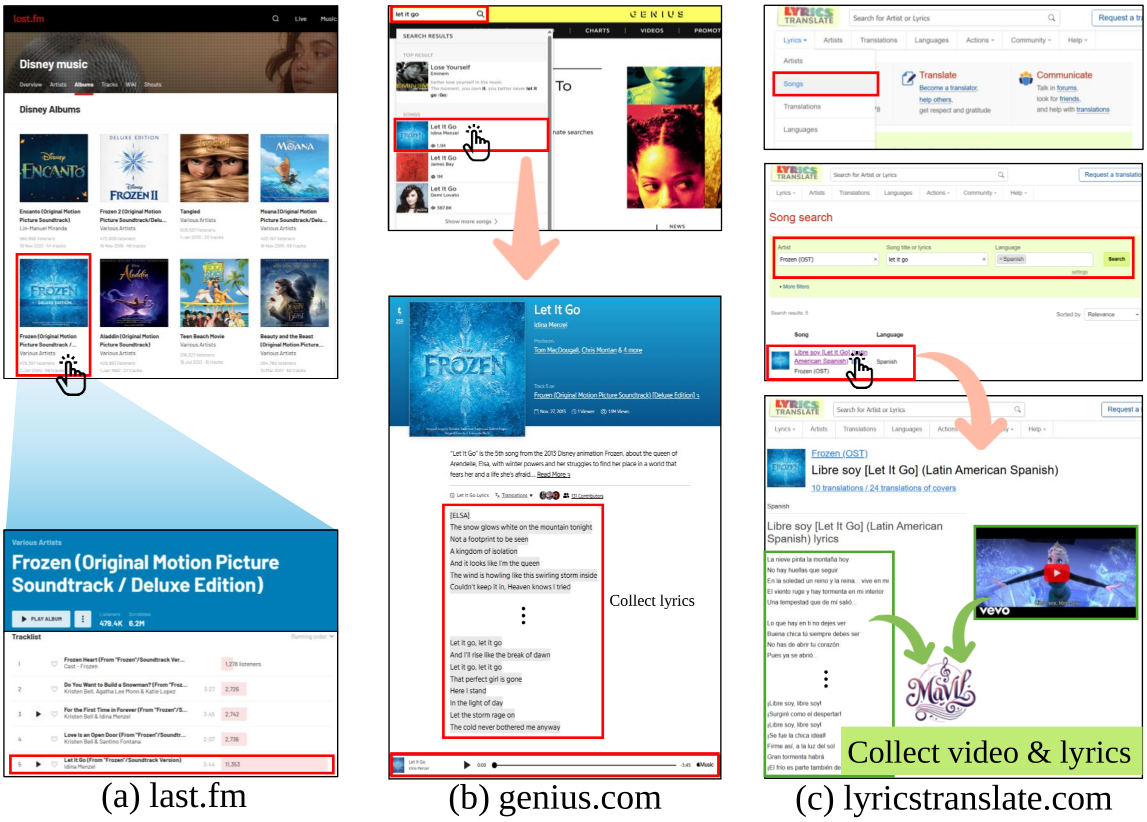

Figure 8: **Web crawling process for MAVL dataset collection.** Our collection pipeline proceeds in three steps—(a) → (b) → (c)—each corresponding to a specific website used in the process. This example is from the Spanish version of the OST "Let It Go" from Frozen, produced by Disney.


```
{
 "gael_garcía_bernal,_gabriella_flores_&_libertad_garcía_fonzi_remember_me_(lullaby)": {
  "youtube_url": {
   "US": "https://www.youtube.com/watch?v=KP_XkN2v7OM",
   "KR": "https://www.youtube.com/watch?v=iBTX6VkU2Gc", ...
  },
  "lyrics_url": {
   "US": "https://genius.com/Gael-garcia-bernal-gabriella-flores-and-libertad-garcia-
fonzi-remember-me-lullaby-lyrics",
   "KR": "https://lyricstranslate.com/en/coco-ost-gieoghae-jweo-
%EA%B8%B0%EC%96%B5%ED%95%B4-%EC%A4%98-remember-me-lyrics.html", ...
  },
  "lyrics": [
   [
    {
     "US": {
      "text": "Remember me, though I have to say goodbye",
      "syllable_count": 11,
      "ipa": "ɹɪmɛmbɹ̩ mi ðow aj hæv tə sej gʊdbaj",
      "line_number": 0,
      "start": 2.4,
      "end": 7.88
     },
     "KR": {
      "text": "기억해 줘 지금 떠나가지만",
      "syllable_count": 11,
      "ipa": "kiʌkʰɛ t͡ɕwʌ t͡ɕigɯm t͈ʌnagadʑiman",
      "line_number": 0,
      "start": 3.84,
      "end": 11.1
     }, ...
    }, ...
   ], ...
  ]
 }
}
```


Figure 9: **MAVL Dataset format.** This is an example of an annotation for a single song. We provide the MAVL dataset in JSON format. This example is “Remember me” from “COCO” by Disney.

model to new languages would require substantial effort in curating new aligned datasets and retraining, making the system inflexible to evolving linguistic needs.

To address these constraints, we leverage **Gemini 2.0 Flash** (DeepMind, 2024), a closed-source MLLM capable of processing audio and video inputs for all six languages. By employing Gemini 2.0 Flash, we bypass the need to develop and train a new model on a highly specialized, low-resource task. Instead, we explore whether prompt-based techniques alone can effectively solve the lyrics translation problem, even for content that requires multimodal understanding.

Our approach centers on adapting Gemini 2.0 Flash via *prompt tuning* rather than extensive finetuning. Specifically, we propose a *Multimodal Chain-of-Thought* pipeline that augments standard chain-of-thought reasoning with additional cues derived from audio and video data. This design allows the model to incorporate contextual information from multiple modalities, which is crucial for translation tasks involving music, animation clips, and other audiovisual elements.

F Prompts for Lyrics Translation

We provide the prompt used by the `SylAVL-CoT` model in Table 15. The prompt for syllable-constrained lyrics translation is in Table 16. The prompt for text-only `SylAVL-CoT` prompt used in ablation Table 6 is in Table 17. We have not included the specific prompts for audio-only and video-only `SylAVL-CoT` ablations, as these are derived directly by combining elements from the text-only and the complete audiovisual `SylAVL-CoT` prompts.

G Sample Reasoning Process for Lyrics Translation

Table 18, 19 shows the reasoning process for Figure 1, 3. Also, Table 20 is the reasoning process example for Appendix C.4.

H Implementation Details

To evaluate and analyze multilingual lyrics, we conducted preprocessing for multilingual data. Since preprocessing must be tailored to the linguistic characteristics of each language, we utilized various libraries, as shown in Table 21. After this process, to account for syllables in numbers, we used num2words [7] to convert numerical values into words, followed by IPA transcription (Mortensen et al., 2018) and syllable counting.

I Experiment Details

Computational Resources

Translating lyrics with Qwen-72B took up to 24 hours, when using 16 NVIDIA A6000s.
the average number of API calls to translate all lines in the dataset is approximately 20,000, which costs about $400K total for Gemini 2.0 Flash.
For all language cross-lingual translation, the number of API calls were about 40,000.

For the generation settings used in Gemini and Qwen, please refer to Table 22 and Table 23, respectively.

[7]https://github.com/savoirfairelinux/num2words

```
You are a professional song translator with expertise in preserving musicality. Translate the following
song lyrics from {source_lang} to {target_lang}.
Please perform the following steps:
1.  Identify the Core Lyric and Perform Syllable Segmentation
- You are given both an audio clip (which may include additional lyrics before or after) and the
corresponding lyric text for a specific scene.
- Use the provided lyric text to determine the exact lyric line you need to process, and disregard
any extraneous audio content that is not part of the given text.
- Listen carefully to the provided audio to capture the natural rhythm, pronunciation, and any important
phonetic, musical, or syllabic characteristics of the identified lyric line. If such notable features
exist, provide a brief explanation of them before proceeding with the translation.
- You will be also given the real syllable count of the original lyric.
- Break down the determined lyric line into its constituent syllables based on the audio's natural
breaks and real syllable count.
- Ensure that each syllable is logically segmented according to the pronunciation. And write it down
in the following format:
Example:
- If the lyric is: "Three months of winter coolness and awesome holidays"
A correct segmentation might be:
   ["Three", "months", "of", "win", "ter", "cool", "ness"]
This original syllable list has 7 syllables.

2.  Generate the Target Language Translation Syllable List Utilizing the Video Information
- Translate the meaning of the lyric naturally and idiomatically into the target language
({target_lang}).
- Review the provided video context and generate the description of the video to understand the
intended mood, imagery, and cultural nuances of the original lyric.
- If specific visuals or cultural elements appear, choose the most context-appropriate term in
{target_lang} to convey the intended meaning.
- Strive to maintain or approximate any rhymes present in the original {source_lang} lyric. You may
modify the literal meaning if it helps preserve rhyme and overall musicality.
- If matching the exact syllable count is too restrictive while trying to keep it a single sentence,
you are permitted to paraphrase more aggressively so that the translation remains fluid and coherent.
- Generate the target language translation syllable list while preserving the original syllable count
whenever possible. Use the audio information to synchronize this syllable list.
- Write down the target language translation syllable list in the following format:
Example:
- Original syllable list: ["Three", "months", "of", "win", "ter", "cool", "ness"]
- Translation syllable list: ["세", "달", "의", "겨", "울", "추", "위"]
The target syllable list has 7 syllables, maintaining the count.

3.  Iterate and Refine the Translation
- After generating the initial translation, check for syllable count, natural flow, rhythm, rhyme, and
meaning in context with the video and audio.
- If the lyric was originally a single sentence, confirm that your translation remains one smooth,
complete sentence in {target_lang}.
- If maintaining the syllable count as a single sentence proves too challenging, continue to refine
your phrasing by paraphrasing, reordering words, or making minor adjustments to meaning. Ensure that
any changes preserve naturalness and singability.
- Repeat this process until the translation feels culturally appropriate, synchronized with the
original audio, and linguistically smooth in {target_lang}.
- Before generating the final translation, please check the syllable count and the translation syllable
list.
- Write down each refined translation in the syllable list format.

4.  Generate the Final Translation
- After ensuring the translation feels natural and maintains the desired structure (single sentence
if the original lyric was one, etc.), use the final syllable list to form the completed translation.
- The final translation should include natural spacing as is customary in the target language. For
languages that do not typically use spacing (e.g., Chinese), do not insert additional spaces.
- Output the final result as a single JSON in the following format:
{"translation": "final translation text"}
Example:
{"translation": "세 달의 겨울 추위"}
   Now, please translate the following {source_lang} lyrics into {target_lang} while fully complying
with the above instructions.
Real Syllable Count: {syllable_count}
Original Lyrics: {source_text}
```

Table 15: **Detailed Prompt for Chain-of-Thought Lyrics Translation**

```
You are a professional song translator with expertise in preserving musicality. Translate the following song lyrics from {source_lang} to {target_lang}.

- Read carefully to the provided lyrics to capture the natural rhythm, pronunciation, and any important phonetic, musical, or syllabic characteristics of the identified lyric line.
- You will be also given the real syllable count of the original lyric.
- Match the syllable count of the original lyric as closely as possible.
- Translate the meaning of the lyric naturally and idiomatically into the target language ({target_lang}).
- If specific cultural elements appear, choose the most context-appropriate term in {target_lang} to convey the intended meaning.
- If matching the exact syllable count is too restrictive while trying to keep it a single sentence, you are permitted to paraphrase more aggressively so that the translation remains fluid and coherent.
- Output the final result as a single JSON in the following format:
“‘json
{ "translation": "final translation text" }
“‘

Now, please translate the following {source_lang} lyrics into {target_lang} while fully complying with the above instructions.

Real Syllable Count: {syllable_count}
Original Lyrics:
{source_text}
```

Table 16: **Detailed Prompt for Syllable-Constraint Lyrics Translation**

```
You are a professional song translator with expertise in preserving musicality. Translate the following song lyrics from source_lang to target_lang.
Please perform the following steps:

1. Identify the Core Lyric and Perform Syllable Segmentation
- Read carefully to the provided lyrics to capture the natural rhythm, pronunciation, and any important phonetic, musical, or syllabic characteristics of the identified lyric line.
- You will be also given the real syllable count of the original lyric.
- Important: You must create a syllable list that matches the syllable count of the original lyric.
- Break down the determined lyric line into its constituent syllables based on real syllable count.
- Ensure that each syllable is logically segmented according to the pronunciation. And write it down in the following format: - Example:
- If the lyric is:
‘"Three months of winter coolness and awesome holidays"‘
- A correct segmentation might be:
‘["Three", "months", "of", "win", "ter", "cool", "ness"]‘
- This original syllable list has 7 syllables.

2. Generate the Target Language Translation Syllable List
- Translate the meaning of the lyric naturally and idiomatically into the target language ({target_lang}).
- If specific cultural elements appear, choose the most context-appropriate term in target_lang to convey the intended meaning.
- Strive to maintain or approximate any rhymes present in the original source_lang lyric. You may modify the literal meaning if it helps preserve rhyme and overall musicality.
- If matching the exact syllable count is too restrictive while trying to keep it a single sentence, you are permitted to paraphrase more aggressively so that the translation remains fluid and coherent.
- Generate the target language translation syllable list while preserving the original syllable count whenever possible.
- Write down the target language translation syllable list in the following format:
- Example:
- Original syllable list: ‘["Three", "months", "of", "win", "ter", "cool", "ness"]‘
- Translation syllable list: ‘["세", "달", "의", "겨", "울", "추", "위"]‘
- Target syllable list has 7 syllables, maintaining the count.

3. Iterate and Refine the Translation
- After generating the initial translation, check for syllable count, natural flow, rhythm, rhyme, and meaning in context.
- If the lyric was originally a single sentence, confirm that your translation remains one smooth, complete sentence in {target_lang}.
- If maintaining the syllable count as a single sentence proves too challenging, continue to refine your phrasing by paraphrasing, reordering words, or making minor adjustments to meaning. Ensure that any changes preserve naturalness and singability.
- Repeat this process until the translation feels culturally appropriate, and linguistically smooth in {target_lang}.
- Before generating the final translation, please check the syllable count and the translation syllable list.
- Write down each refined translation in the syllable list format.

4. Generate the Final Translation
- After ensuring the translation feels natural and maintains the desired structure (single sentence if the original lyric was one, etc.), use the final syllable list to form the completed translation.
- The final translation should include natural spacing as is customary in the target language. For languages that do not typically use spacing (e.g., Chinese), do not insert additional spaces.
- Output the final result as a single JSON in the following format:
“‘json
"translation": "final translation text"
“‘
- Example:
“‘json
"translation": "세 달의 겨울 추위"
“‘
```

Table 17: **Detailed Prompt for Text-Only Chain-of-Thought Lyrics Translation**

```
1. Identify the Core Lyric and Perform Syllable Segmentation
- The core lyric is: "And there's a butterfly".
- Audio analysis indicates a natural flow with emphasis on "butterfly".
- The original syllable count is 6.
- Syllable segmentation: '["And", "there's", "a", "but", "ter", "fly"]'

2. Generate the Target Language Translation Syllable List Utilizing the Video Information
- Video context: The scene shows a colorful, cheerful environment with a pink Troll, running happily.
A butterfly flits around. The mood is light and joyful.
- Translation considerations: We need a Korean translation that fits this happy, whimsical context
and maintains the 6-syllable count. We can use a common Korean word for butterfly, "나비" (nabi).
- Initial translation: "나비 가 있 어 요" (Na-bi ga it-eo-yo) - This translates to "There's a
butterfly."
- Syllable list: '["나", "비", "가", "있", "어", "요"]'

3. Iterate and Refine the Translation
- The initial translation "나비 가 있어요" is grammatically correct and fits the context. The syllable
count is also correct at 6.
- We can consider other options to see if we can get a more singable result, but this is already quite
good.
- Let's try "나비 가 날 아 와" (Na-bi ga nal a wa) - "A butterfly comes flying"
- Syllable list: '["나", "비", "가", "날", "아", "와"]'
- This also has 6 syllables and fits the scene well. It emphasizes the movement of the butterfly,
which is visually present.

4. Generate the Final Translation
- I think "나비 가 날 아 와" is slightly better because it has a sense of movement.
- Final Syllable List: '["나", "비", "가", "날", "아", "와"]'
- Final Translation: 나비가 날아와

"'json
{"translation": "나비가 날아와"}
"'
```

Table 18: **Sample Reasoning Process for Chain-of-Thought Lyrics Translation**

1. Identify the Core Lyric and Perform Syllable Segmentation
I begin with the original lyric “Remember me, don’t let it make you cry” and note the natural break after “me.” I segment the lyric into its constituent syllables as follows:
‘["Re", "mem", "ber", "me,", "don’t", "let", "it", "make", "you", "cry"]‘
This gives a total of 10 syllables.

2. Generate the Target Language Translation Syllable List
The video shows a young boy singing to his elderly grandmother, who looks sad.
Given the context, the translation should be comforting and gentle. I will aim for a translation that conveys the same emotion and fits the rhythm of the song.
My initial translation is:
‘["날", "기", "억", "해", "줘", "울", "지", "는", "마"]‘
This translation maintains the sentiment and cultural context.
3. Iterate and Refine the Translation Upon reviewing the segmentation ["날", "기", "억", "해", "줘", "울", "지", "는", "마"], several issues become apparent:
Upon reviewing the initial segmentation ["날", "기", "억", "해", "줘", "울", "지", "는", "마"], the primary issue is the syllable count.
The list contains 9 syllables, not 10 as originally noted. This mismatch is crucial for accurately fitting the lyric to the song’s original melody and rhythm, which is built on 10 syllables.
Furthermore, the original lyric has a distinct 4+6 syllable structure with a natural break after the 4th syllable ("me,"). The initial translation has a 5+4 structure (["날", "기", "억", "해", "줘"] as the first part and ["울", "지", "는", "마"] as the second), causing the potential break point to fall after the 5th syllable instead of the 4th.
This misalignment in both total syllable count and internal structure will disrupt the timing and flow of the lyric when sung to the original music, making it difficult to match the notes and rhythm.

Second attempt (adjusting for syllable count and naturalness):
‘["날", "잊", "지", "마", "슬", "퍼", "하", "지", "는", "마"]‘
The second attempt seems more natural and maintains the 10-syllable count.
Let’s check the meaning and flow. It translates roughly to "Don’t forget me, don’t be sad". This captures the essence of the original lyric.
4. Generate the Final Translation The final syllable list is:
‘["날", "잊", "지", "마", "슬", "퍼", "하", "지", "는", "마"]‘

“‘json
{"translation": "날 잊지 마 슬퍼하지는 마"}
“‘

Table 19: **Sample Reasoning Process for Chain-of-Thought Lyrics Translation**

```
1. Identify the Core Lyric and Perform Syllable Segmentation
The original lyric is “기억해 줘 지금 떠나가지만”.
I segment the lyric into its constituent syllables as follows:
‘["기", "억", "해", "줘", "지", "금", "떠", "나", "가", "지", "만"]‘
This gives a total of 11 syllables.

2. Generate the Target Language Translation Syllable List
The video shows a young boy singing to an elderly woman in a wheelchair. The boy is trying to cheer
her up with the song.
Given the context, I need to translate the line to convey a sense of longing and remembrance as the
singer departs. I aim to match the syllable count and create a natural flow.
My initial translation is:
‘["Re", "mem", "ber", "me", "though", "I", "have", "to", "leave", "you", "now"]‘

3. Iterate and Refine the Translation
The initial translation has 11 syllables, which matches the original. The phrase "remember me"
captures the essence of the original lyric "기억해 줘". The phrase "though I have to leave you now"
translates "지금 떠나가지만" effectively, while maintaining a natural rhythm.

4. Generate the Final Translation
The translation meets all the required conditions: natural flow, correct syllable count, and cultural
appropriateness.
The final English translation is:
"Remember me though I have to leave you now"

“‘json
{
"translation": "Remember me though I have to leave you now"
}
“‘
```

Table 20: **Sample Reasoning Process for Chain-of-Thought Lyrics Translation From Korean to English**

	Syllable	IPA
EN	Syllabifier [8]	epitran [9]
ES	count-syllables-in-spanish [10]	epitran
FR	syllabify-fr [11]	epitran
KO	length of text	epitran
JA	japanese-mora-counter [12]	epitran

Table 21: **Libraries used for syllable counting and IPA conversion across languages.**

Parameter	**Value**
`temperature`	0.6
`top_p`	0.95
`top_k`	40
`max_output_tokens`	8192
`response_mime_type`	`text/plain`

Table 22: **Gemini generation configuration.**

Parameter	**Value**
`temperature`	0.7
`top_p`	0.8
`max_tokens`	4096
`presence_penalty`	1.05

Table 23: **Qwen generation configuration.**