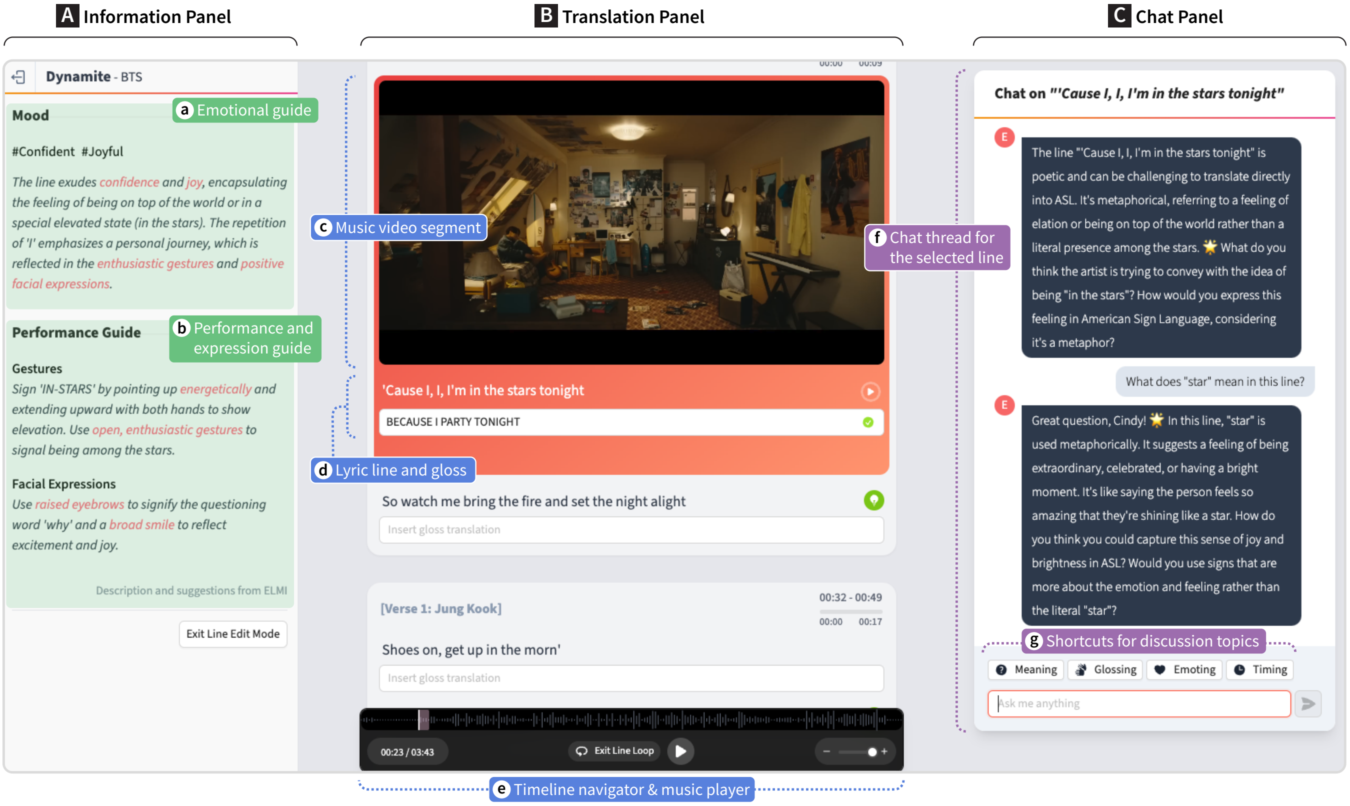
ELMI: Interactive and Intelligent Sign Language Translation of Lyrics for Song Signing

Suhyeon Yoo*
University of Toronto
Toronto, Canada
suhyeon.yoo@mail.utoronto.ca

Khai N. Truong
University of Toronto
Toronto, Canada
khai@cs.toronto.edu

Young-Ho Kim
NAVER AI Lab
Republic of Korea
yghokim@younghokim.net



#Confident #Joyful

The line exudes confidence and joy, encapsulating the feeling of being on top of the world or in a special elevated state (in the stars). The repetition of 'I' emphasizes a personal journey, which is reflected in the enthusiastic gestures and positive facial expressions.

Gestures: Sign 'IN-STARS' by pointing up energetically and extending upward with both hands to show elevation. Use open, enthusiastic gestures to signal being among the stars.

Facial Expressions: Use raised eyebrows to signify the questioning word 'why' and a broad smile to reflect excitement and joy.

'Cause I, I, I'm in the stars tonight

BECAUSE I PARTY TONIGHT

So watch me bring the fire and set the night alight

Shoes on, get up in the morn'

Chat on "'Cause I, I, I'm in the stars tonight"

E: The line "'Cause I, I, I'm in the stars tonight" is poetic and can be challenging to translate directly into ASL. It's metaphorical, referring to a feeling of elation or being on top of the world rather than a literal presence among the stars. 🌟 What do you think the artist is trying to convey with the idea of being "in the stars"? How would you express this feeling in American Sign Language, considering it's a metaphor?

What does "star" mean in this line?

E: Great question, Cindy! 🌟 In this line, "star" is used metaphorically. It suggests a feeling of being extraordinary, celebrated, or having a bright moment. It's like saying the person feels so amazing that they're shining like a star. How do you think you could capture this sense of joy and brightness in ASL? Would you use signs that are more about the emotion and feeling rather than the literal "star"?

Figure 1: ELMI's translation interface, supporting line-by-line translation of lyrics into sign language. The Information Panel A provides background information about the song or AI-generated description on mood and performance guide for the selected lyric line. The emotional guide ⓐ suggests hashtags that capture the mood of each line, while the performance and expression guide ⓑ offers recommendations for bodily and facial expressions to effectively convey these emotions. On the Translation Panel B, the user can write gloss (text representation of sign language) for the selected lyric line ⓓ, looping the music video segment ⓒ. The Chat Panel C provides a discussion channel with a large language model-driven chatbot regarding the selected line ⓕ. For the paper figures, we used the official music video of BTS's Dynamite, publicly available at Youtube (https://www.youtube.com/watch?v=gdZLi9oWNZg).

Abstract

d/Deaf and hearing song-signers have become prevalent across video-sharing platforms, but translating songs into sign language remains cumbersome and inaccessible. Our formative study revealed the challenges song-signers face, including semantic, syntactic, expressive, and rhythmic considerations in translations. We present ELMI, an accessible song-signing tool that assists in translating lyrics into sign language. ELMI enables users to edit glosses line-by-line, with real-time synced lyric and music video snippets. Users can

---

[*] Suhyeon Yoo conducted this work as a research intern at NAVER AI Lab.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI '25, April 26-May 1, 2025, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713973

also chat with a large language model-driven AI to discuss meaning, glossing, emoting, and timing. Through an exploratory study with 13 song-signers, we examined how ELMI facilitates their workflows and how song-signers leverage and receive an LLM-driven chat for translation. Participants successfully adopted ELMI to song-signing, with active discussions throughout. They also reported improved confidence and independence in their translations, finding ELMI encouraging, constructive, and informative. We discuss research and design implications for accessible and culturally sensitive song-signing translation tools.

CCS Concepts

• **Human-centered computing → Accessibility systems and tools**; **Empirical studies in accessibility**.

Keywords

Song Signing, Deaf Music, Lyrics Translation, Large Language Model, Human-AI Collaboration, Creativity Support

ACM Reference Format:
Suhyeon Yoo, Khai N. Truong, and Young-Ho Kim. 2025. ELMI: Interactive and Intelligent Sign Language Translation of Lyrics for Song Signing. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3706598.3713973

1 Introduction

Song-signing[1] is a performing art, encompassing not only translating spoken lyrics to sign language but also conveying musical elements through facial expressions and bodily movements [7, 35]. d/Deaf[2] and hearing song-signers perform in visually dynamic forms, blending the rhythm and emotion of songs with the physical expressiveness of sign language [23]. Translation quality is crucial, as it enhances the message of the song by incorporating emotional tone, timing, and cultural nuances, going beyond the literal meaning of the words [7, 16, 35]. This requires a deep understanding of both spoken and sign languages, ensuring that the final signed performance is both accurate and resonant [64, 74].

However, translating song lyrics into signs presents challenges. Often, sign language may not have the equivalent vocabulary for some terms from the spoken language, necessitating the use of fingerspelling or creative alternatives [74]. Furthermore, since the syntax and modality of two languages are distinct, the *glosses*[3] should be adjusted to fit the rhythm of the music and the signing space performing sign language [64, 74]. While past research has explored ways to make music more accessible for d/Deaf individuals, much of this work has focused on music perception [30, 38] and creation [46, 51] rather than performance. Research on the artistic interpretation of sign language remains sparse, and prior work predominantly focused on communication accessibility [58]. Only a few projects have recently begun to understand song-signing [74]. For example, Yoo *et al.* identified tensions between d/Deaf and hearing song-signers and the different barriers they encounter in the process of song-signing. Still, little effort has been put into designing accessible systems for song-signing.

In this work, we aim to design a system that supports translation of song lyrics and gloss creation for song-signing in a more accessible manner. To better understand the current song-signing translation process and challenges, we conducted formative interviews with eight song-signers (five d/Deaf and three hearing). The study revealed four major challenges of song-signing in the translation process: semantic translation (*e.g.*, understanding the meaning of the lyrics), syntactic translation (*e.g.*, selecting appropriate signs) [4], expressive translation (*e.g.*, conveying emotion), and rhythmic translation (*e.g.*, aligning the timing and length of signs with the music). As a result, many song-signers rely on external resources such as social media videos and discussions with other interpreters to improve their work.

To address these challenges, we designed and developed **ELMI** (Explore Lyrics and Music Interactively; Figure 1), a web-based accessible song-signing tool that facilitates lyric-sign translation. ELMI features a line-by-line focus, allowing the user to input gloss for each line and play corresponding video segments. For a d/Deaf user, ELMI offers rich visual feedback to convey music, such as a music video player with a karaoke-style timing indicator. Inspired by the generative capabilities of large language models (LLMs), ELMI leverages LLMs to provide various AI-driven guidance. For example, to aid in familiarization with the song, ELMI provides AI-generated annotations of the lyric lines, such as emotional tone and performance guidance. For each lyric line, the user can chat with ELMI to discuss key topics of gloss creation, including meaning, glossing, emotion, and timing.

We evaluated ELMI through an exploratory study with 13 song-signers (6 d/Deaf, and 7 hearing), who used the tool to create glosses for two songs. ELMI facilitated the translation process by producing nuanced and contextually accurate glosses. Participants reported increased confidence and independence in their translation process through their discussions with ELMI. Both d/Deaf and hearing participants found glossing and meaning-related discussions useful while hearing favored emoting, and d/Deaf valued timing assistance. ELMI streamlined decision-making, providing all required resources in one place. Participants perceived ELMI as encouraging, critical, and informative in giving them full control and ownership over their glosses. All participants expressed interest in incorporating ELMI into their workflow, as it helped them structure their process. The key contributions of this work are threefold:

---

[1] Also known as sign singing, signed song, sign singing, or karaoke signing [16]. Song-signing involves translations/interpretations of pre-existing songs into a signed language, original pieces composed in sign language, or arrangements of signs to specific beats known as percussion signing (*e.g.*, "Bison Song" by Dorothy Miles) [7, 35]. The signed songs can be presented as recorded videos (*e.g.*, YouTube performances) or as live presentations (*e.g.*, concerts or stage performances) [7, 35].

[2] In research, the acronym DHH (Deaf and Hard of Hearing) is often used to refer to individuals with varying degrees of hearing loss [27]. However, in this study, we use "d/Deaf" to emphasize cultural identity. "deaf" (with a lowercase 'd') refers to the physical condition of hearing loss. These individuals may use hearing aids, cochlear implants, or rely on lip reading. On the other hand, "Deaf" (with an uppercase 'D') refers to individuals who identify as part of the Deaf community, share a common language (such as ASL), and embrace Deaf culture and identity [63].

[3] Gloss refers to a written representation of a signed language, using words from a spoken language (like English) to indicate the meaning of individual signs. Glossing helps translate the structure and meaning of a signed language but does not fully capture nuances such as facial expressions, which are crucial elements of communication [68].

[4] In this work, we use the term *syntactic translation* to indicate the process of translating lyrics into glosses while ensuring that the sign order aligns with ASL grammar and maintains the musical timing [74].

(1) The design and implementation of ELMI, an AI-infused song-signing tool that supports translating lyrics into sign language. The source code of ELMI is publicly available at `https://naver-ai.github.io/elmi`.

(2) The empirical understanding of how song-signers perceive and interact with ELMI, creatively producing glossing through discussion with the tool.

(3) Implications for future research and systems on acceptability and accountability for artistic sign language translation, highlighting the need to respect cultural contexts and maintain artistic integrity.

2 Related Work

In this section, we cover related work in three parts: (1) song-signing, (2) technologies for automated sign language translation, and (3) lyric translation technologies.

2.1 Song-signing

Song-signing is an art form rooted in Deaf culture [35]. Song-signers convey not only the lyrics but also capture musical elements such as rhythm, emotion, and instruments through facial expressions and bodily gestures [74]. Song-singing performance is crucial in making music accessible to d/Deaf individuals by providing an augmented visual representation [74]. It has also proved to be more meaningful and authentic to those fluent in sign language, offering a genuine and enjoyable experience [17].

The core activity of song-signing—trans-languaging from spoken language to sign language—requires a high degree of creativity, encompassing the translation of lyrics or the composition of original songs in sign language, where performers manipulate signs and signing space to convey musical features [35]. This practice involves complex interactions between music, lyrics, and sign language, requiring artists to navigate across languages and modalities [64]. Yoo *et al.* proposed three layers in song-signing translation: technical translation, understanding the lyrics and matching gloss with the music; artistic translation, editing the gloss to convey non-lyrical elements like pitch, rhythm, and volume; and cultural translation, appreciating of Deaf culture and language [74]

Recent studies on song-signing have identified several challenges, including the difficulty of translating lyrics into sign language, conveying musical features, and addressing diverse linguistic repertoires [35, 74]. Song-signing involves navigating between visual and auditory channels and integrating music, lyrics, and sign language [64]. One significant challenge is that many song lyrics carry symbolic meanings rather than straightforward messages. When translated into sign language, the interpretation is shaped by the translator, which can lead to a disconnect between the original intent of the writer and the signed interpretation [17]. Additionally, song-signing can appear weak if the translator struggles with timing, particularly when trying to produce musical American Sign Language (ASL) within the time constraints of English songs [17]. Although these challenges underscore the need for culturally responsive approaches to music accessibility [57, 74], research on how technology can support song-signing translation remains limited.

2.2 Automated Sign Language Translation

Sign language is the primary mode of communication for d/Deaf individuals [59], and research in this area is crucial for facilitating communication between Deaf communities and others [50]. As one dominant line of efforts that bridge the gap between spoken and sign languages, ample research in machine learning and accessibility has presented automated models that convert signed sentences into spoken or written language and vice versa [19, 33, 44, 50]. As sign language is carried over bodily movements, video recognition or generation is required for ideal end-to-end translation [33]. To make the task straightforward, gloss text has often been used as a representation medium of sign language, and a majority of machine learning approaches tackled gloss-spoken language translation [19, 44]. For example, researchers proposed gloss-based neural machine translation models (*e.g.*, [1, 13, 14, 42]), Recognizing the loss of partial linguistic cues when using gloss as a representation of sign language [19, 43], recent work leveraged large language models (LLMs) [47] to implement gloss-free sign language translation in combination with vision models [25, 73].

Despite the growing body of automated sign language translation technologies, existing models focus on communication and mapping the meanings, trained on datasets in general contexts such as word-level sign lexicons (*e.g.*, [6, 32]), news and weather forecast (*e.g.*, [12]), and daily conversations (*e.g.*, [70]). In other words, they are not designed with lyric translation in mind, which critically involves considerations of timing and poetic interpretation [74]. Hence, lyric-to-sign-language translation should incorporate users in the loop, rather than being carried over a fully-automated approach. In this work, we support lyric translation driven by song-signers, while facilitating the translation process with AI-driven guidance and discussions.

2.3 Lyric Translation and Tools

Emotion and timing are crucial in translating lyrics due to their significant impact on the overall musical experience. Translators must navigate the complex interplay between language, music, and cultural nuances to maintain the emotional resonance of the original song [69]. This translation process requires careful consideration of both implicit and explicit information, as well as the emotional content of the source text [29]. Franzon describes the strategic choices made by translators and lyricists in achieving 'singability,' the musico-verbal fit of text to music—by balancing prosodic, poetic, and semantic layers [24]. The emotional intelligence of translators also plays a pivotal role in their decision-making and problem-solving during translation: understanding of linguistic, cultural, and emotional subtleties to effectively bridge linguistic and cultural boundaries in music [29, 69].

New tools for lyric translation and interpretation have been introduced. Statistical Machine Translation was applied to an automated lyric annotation system to clarify complex terminology and abstract concepts using the Genius dataset [65]. Semi-automatic lyric generation using context-free grammar has shown promise, with generated lyrics sometimes indistinguishable from those written by humans [53]. Recent research has explored the potential of large language models (LLMs) in supporting lyric interpretation and song understanding. For instance, BART-fusion combines a pre-trained language model with an audio encoder to generate lyric interpretations, showing improved performance with the integration of audio information [76]. These studies demonstrate the versatility of LLMs in understanding and generating music-related content, highlighting their potential to enhance interpretability across various domains.

Despite these technological advancements, the translation of lyrics continues to pose significant challenges. Translators must capture stylistic elements, cultural references, and semantic nuances while maintaining fluency and creativity [2]. The ambiguous nature of lyrics, along with the use of slang, jargon, and the need to preserve rhyme and repetition, adds to the complexity of this task [3]. Moreover, translators face difficulties related to rhythm, syllable count, vocal burden, and the need to adapt cultural nuances across languages [3]. Cultural differences often necessitate adaptations and retranslations to accurately convey meaning. Translating idioms within lyrics, in particular, requires a deep understanding of both the source and target languages to ensure the intended meanings are preserved [2]. Addressing these challenges requires interdisciplinary approaches that combine linguistics, artificial intelligence, and cultural studies. This work extends the body of research on lyric translation within the context of song-signing, focusing on artistic and cultural dimensions in the translation process.

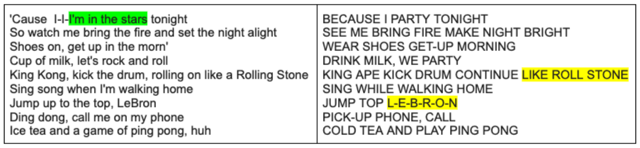
3 Formative Study

To understand the current translation process and challenges in song-signing, as well as to explore the applicability of LLM on song-signing translation, we conducted semi-structured interviews with song-signers. The study protocol was approved by the university research ethics board.

3.1 Methods

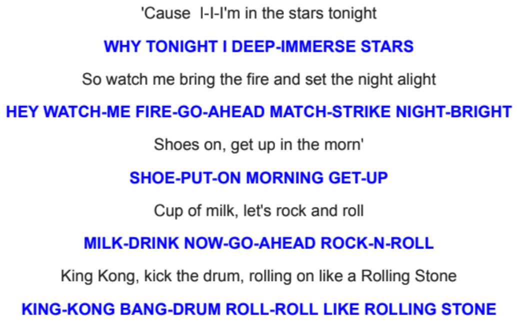
Participants. We recruited eight song-signers ($S1^{d/D}$ – $S8^{Hg}$; five d/Deaf and three hearing in order. See Appendix A.1 for detailed demographic information.) comprising five males and three females, with ages ranging from 25 to 67 years old. Our inclusion criteria were song-signers who are 18 years or older, proficient in speaking and writing English, and have experience in song-signing within the past year. Participants were recruited through emails, social media platforms (e.g., Facebook, YouTube), and the snowball sampling method. The majority of participants ($N = 5$) used American Sign Language (ASL) at expert and native levels, while others also utilized Pidgin Signed English (PSE). On average, participants have been involved in song-signing for 9.3 years, spanning from 4 to 17 years. While some participants used tools such as dictionaries for translation, the use of LLM-driven tools like ChatGPT was minimal, with most participants relying on personal knowledge and community resources.

Procedure. We invited each participant to Zoom sessions which lasted between 50 to 70 minutes. We utilized Zoom chats, closed captioning, and sign language interpreters for d/Deaf participants who preferred those options. Participants' signed responses were interpreted in real-time by certified ASL interpreters, whose spoken translations were then transcribed into English. The interviews covered (1) the current work process and practice method for translating English lyrics to sign language, (2) factors and considerations influencing translation, and (3) challenges and workarounds in translation. To collect participants' preliminary feedback on using LLMs in song-signing, we asked them to try ChatGPT [48] while translating a snippet of the example song "Dynamite" by BTS (See Figure 2). During this task, participants were encouraged to ask ChatGPT any open-ended questions that came to their mind regarding various aspects of the translation process. Participants engaged ChatGPT in diverse ways, such as understanding the context of the song, interpreting specific lyrics, seeking confirmation on glossing and interpretation, finding videos of performances or song-signings, and providing images or videos of example signs.



'Cause I-I-I'm in the stars tonight	BECAUSE I PARTY TONIGHT
So watch me bring the fire and set the night alight	SEE ME BRING FIRE MAKE NIGHT BRIGHT
Shoes on, get up in the morn'	WEAR SHOES GET-UP MORNING
Cup of milk, let's rock and roll	DRINK MILK, WE PARTY
King Kong, kick the drum, rolling on like a Rolling Stone	KING APE KICK DRUM CONTINUE LIKE ROLL STONE
Sing song when I'm walking home	SING WHILE WALKING HOME
Jump up to the top, LeBron	JUMP TOP L-E-B-R-O-N
Ding dong, call me on my phone	PICK-UP PHONE, CALL
Ice tea and a game of ping pong, huh	COLD TEA AND PLAY PING PONG

(a) Excerpt of glosses from $S6^{Hg}$



'Cause I-I-I'm in the stars tonight

WHY TONIGHT I DEEP-IMMERSE STARS

So watch me bring the fire and set the night alight

HEY WATCH-ME FIRE-GO-AHEAD MATCH-STRIKE NIGHT-BRIGHT

Shoes on, get up in the morn'

SHOE-PUT-ON MORNING GET-UP

Cup of milk, let's rock and roll

MILK-DRINK NOW-GO-AHEAD ROCK-N-ROLL

King Kong, kick the drum, rolling on like a Rolling Stone

KING-KONG BANG-DRUM ROLL-ROLL LIKE ROLLING STONE

(b) Excerpt of glosses from $S7^{Hg}$

Figure 2: Example Glossing for "BTS - Dynamite." Song-signers created glosses line-by-line, writing ASL glosses corresponding to ENG lyrics.

Analysis. We recorded the video call sessions and transcribed them for analysis. Using a thematic analysis framework [9, 10], two researchers independently coded the transcripts to identify key ideas and grouped similar codes into categories. Through multiple rounds of discussion, the researchers refined and reviewed themes to ensure they accurately reflected the data and aligned with the research objectives.

3.2 Lyric Translation Practice

3.2.1 Familiarization. Participants begin by listening to the music and watching music videos to understand the context of the song's lyrics. Rather than focusing on a word-for-word translation, song-signers aim for *conceptual accuracy*, ensuring that the essence and message of the song are preserved. During this process, an initial gloss is created that captures the general meaning of the song. The lyrics are then broken down into chunks and eventually "line by line", with each segment translated into sign language, ensuring that each sign fits within the overall narrative of the song. Additionally, participants conducted contextual and cultural research to understand the background and significance of the song: "*I do a little bit of research why the lyric was written, why the artist chose the song, and how they sang it* (S1$^{d/D}$)."

3.2.2 Practice and Refinement. After creating the initial version of the gloss, participants refine it to ensure the smoothness and visual appeal of the translation. S4$^{d/D}$ remarked, "*In practice, I have an idea of what I want to sign. Sometimes I sign with the music and come across certain spots that are very tricky. Then I might go back and reevaluate if there's a simpler way to sign it.*" Artistic elements such as rhyming were emphasized to retain the artistic integrity of the original song and make the performance visually pleasing. Two participants (S2$^{d/D}$, S6^{Hg}) noted the importance of utilizing the five parameters of ASL—handshape, location, movement, palm orientation, and non-manual markers.

Participants commonly used glosses to understand and memorize the signs. Participants also used glossing while performing, often setting up cameras with reference glosses with lyrics positioned nearby for visibility. Half of the participants (N = 4) used online editing tools like Google Docs to create glosses (See Figure 2). However, as song-signers became more experienced, they tended not to use written glosses but to practice directly with the music.

3.3 Challenges of Song-signing Translation

Semantic Translation [Meaning]. Participants pointed out the difficulty of grasping the meaning of the lyrics. This challenge arises from the inherent complexity of the lyrics, including poetic or ambiguous expressions that can have multiple meanings (N = 6). Accurately capturing meaning becomes even more challenging when translating concepts that are culturally or linguistically specific. S8^{Hg} noted how accents or unfamiliar terms can lead to misinterpretations: "*[...] Another challenge is the accent. I might miss something because I don't understand a particular term.*"

Syntactic Translation [Glossing]. Since ASL relies heavily on visual elements to convey meaning, participants noted that it is particularly difficult to translate concepts that are abstract or primarily auditory in nature. For example, three participants mentioned challenges in translating repetitive or simplistic messages in a way that feels both meaningful and visually appealing. Another challenge arises when English phrases do not have direct equivalents in ASL, requiring signers to find alternative signs or creative ways to convey the meaning. Also, acoustic rhymes in spoken language, which depend on sound, do not translate well into sign language.

Expressive Translation [Emoting]. Participants pointed out that beginners or those new to Deaf culture would often struggle with emotional demands in sign language due to their lack of experience and the mismatch between their own personality and the song's mood. S7^{Hg} remarked, "*It's a challenge, especially if you're a person who may be outwardly calm or just don't like to express yourself. Then you really have to work on that.*" Certain songs, particularly those with shifting moods, demand expressive control that can be hard to master. S6^{Hg} shared, "*I'm not the best at controlling my face. It sometimes looks awkward...Why aren't you smiling enough? One challenge is self-consciousness about your face.*"

Rhythmic Translation [Timing]. The discrepancy of modalities of spoken and signed languages imposed challenges, especially for fast-paced songs, as the rapid tempo can strain even the most skilled signers. Moreover, achieving synchronization with the music adds another layer of complexity, particularly for deaf signers. S5$^{d/D}$ described the need for extensive repetition to match the signs with the lyrics: "*It was tough, I had to repeat and repeat until I could be in sync with the music. I would repeat one sentence over 50 times to feel confident to sign as they sing.*" This meticulous process highlights the challenge of aligning signs with the music, a task that demands significant time and effort.

3.4 Common Strategies to Overcome Challenges

Participants commonly looked up external resources (N = 5). Many turned to online ASL dictionaries to find the appropriate signs and expressions, appreciating that the short video clips and accompanying descriptions helped clarify hand positions and sign execution. Participants also relied on search engines like Google, particularly for searching the origins and meanings of songs. Various types of videos—including original music videos, performance/dance videos, and song-signing videos from other song-signers—also served as a reference. All participants mentioned watching these videos helped them interpret the mood of the song and the underlying messages of the lyrics. Participants referred to performance videos to observe facial expressions, energy levels, and choreography, which they then incorporated into their translations. Many participants also watched other song-signing videos to gain ideas and inspiration.

Notably, most participants (N = 6) mentioned that they discussed their work with others during the translation process. They gathered feedback from a diverse group, including interpreters, Deaf individuals, and young people, to refine their interpretations. When consulting hearing individuals, participants often sought advice on visual quality and how well the lyrics matched the music. As for consulting interpreters or Deaf individuals, the focus was on refining the overall translation. Participants also consulted young people to understand contemporary slang and language trends.

3.5 Feedback on Using LLMs for Translation

The majority of participants shared positive views on using GPT in the translation process, recognizing its capabilities extending beyond mere literal translation, as it offers interpretations that capture the deeper meaning and emotional nuances of the original content. Additionally, GPT's recommendations for incorporating non-manual signs were particularly impressive to participants. For example, S6^{Hg} noted: "*My reaction was amazing as far as I can see... It asked me to use the classifier, where I found the power of that.*" Despite its potential, participants also expressed concerns about relying on it for accurate and contextually appropriate translations because ChatGPT sometimes yielded wrong glosses or interpreted the lyrics literally, producing very long glosses.

4 ELMI: Explore Lyrics and Music Interactively

Informed by the formative study, we designed and developed ELMI, a novel web application that aids d/Deaf and hearing individuals in song-signing. Our approach leverages the potential of LLMs to facilitate the glossing process, which our formative study revealed as a critical foundation for song-signing translation. By facilitating

AI-assisted discussions around translation nuances—such as semantic accuracy, syntactic clarity, and expressive interpretation—ELMI supports users in addressing the challenges highlighted by the song-signers in the formative study. In the following, we outline the design rationales for ELMI and describe the details of the interface and its implementation.

4.1 Design Rationale

DR1: Use sufficient visual aids to convey emotion and timing. Music can be appreciated not only by auditory but also by visual channels [20, 66]. Formative study participants also relied on lip-reading from performance videos and utilizing captions to interpret the lyrics. This aligns with prior research, where d/Deaf individuals often lip-read and observe expressions of singers to capture musical elements [11], while captions significantly enhance their comprehension and enjoyment [40]. Visual information is thus essential for conveying musical elements like emotion and timing. Consistent with prior work [74], formative study participants reported using visual aids, including online videos. To address this, we designed ELMI to provide music videos with captions synchronized to the lyrics (*e.g.*,ⓒ in Figure 3), allowing users to fully grasp the nuances and accurately translate them.

DR2: Support discussion about nuances, performance, and translation of lyrics. Participants in formative study commonly engaged in discussions about song interpretation, mirroring how individuals use online forums to learn and share insights [8, 45, 71]. They also recognized the potential of tools like ChatGPT to support their workflow, aligning with recent work on using ChatGPT for sign language translation [49, 61]. Based on these findings, we incorporated an LLM-driven chat feature to ELMI, which allows discussions across four key topics: meaning, glossing, emoting, and timing. We intended to facilitate brainstorming and the exchange of ideas, allowing song-signers to gain different perspectives and supporting users in making artistic decisions.

DR3: Focus translation at a line level. Unlike typical text, song lyrics are structured in lines rather than sentences and paragraphs, and they often feature shorter, repetitive phrasing [15]. We found from the formative study that song-signers typically gloss lyrics line by line. In response, we designed ELMI to facilitate gloss creation, discussion, and playback of the music video on a line-by-line basis. This approach helps users break down the lyrics into manageable sections, allowing them to focus on translating each line, while also accommodating the iterative nature of refining individual lines to achieve the best possible interpretation. We initially considered supporting AI discussions about the general information of the song to help users familiarize themselves with it before translation. However, formative study participants rarely discussed such topics with ChatGPT, and most preferred static materials or performance videos for initial research. As a result, we limited AI discussions to the line-by-line translation.

4.2 User Interface and Interaction Components

ELMI operates in two modes: By default, the system is in the **Global Play** mode, where the user can play the entire video using a conventional video player (ⓔ in Figure 1). In this mode, the **Information Panel** (Figure 1-A) serves as the user's entry point for exploring the story and meaning of the song; it displays essential song details, such as genre, runtime, and behind-the-scenes insights, all sourced from the song description of Genius music database [39]. While the song is being played, the **Translation Panel** (Figure 1-B) automatically scrolls to display the current line.

The user can enter the **Line Loop** mode by selecting a lyric line on the Translation Panel. In this mode, the selected line is highlighted (Figure 1-B, top), and the user can create and refine glosses (ⓓ in Figure 1) with real-time feedback from the chatbot (ⓕ in Figure 1). In this mode, the Information Panel displays the AI-generated analysis results for the line with guidance on emotional expression such as non-manual signals (NMS), including a Mood section (ⓐ in Figure 1) with hashtagged emotions, and a Performance Guide (ⓑ in Figure 1) offering recommendations on gestures and facial expressions (See Figure 1-A).

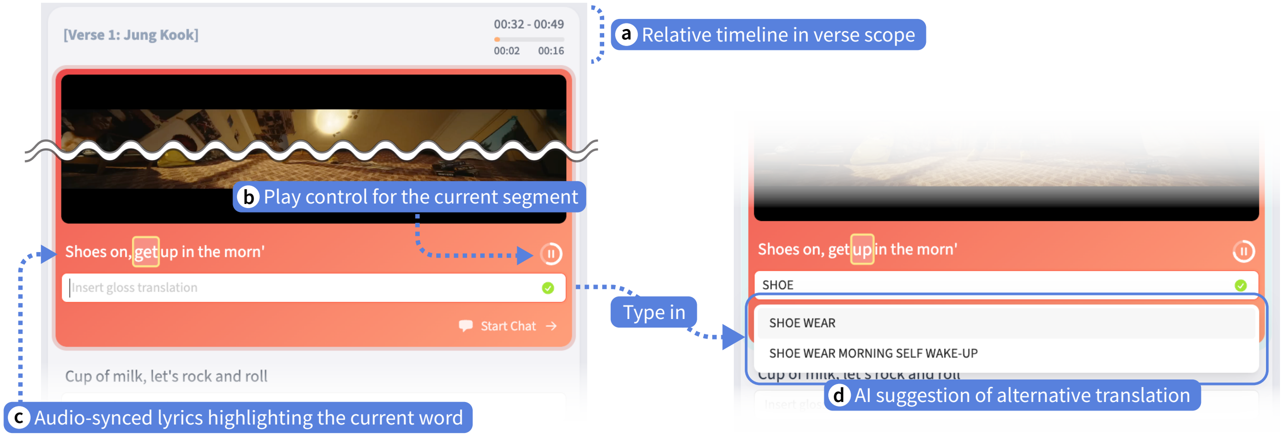
On the Translation Panel, the user can write glosses directly beneath each line of lyrics (See Figure 3). As they type, the system provides real-time suggestions for one or two alternative versions of the current input gloss, which users can choose to adopt (See ⓓ in Figure 3). The user can still play and pause the video, but the current line segment is looped (See ⓑ in Figure 3).

ELMI also supports discussion with an LLM-driven chatbot on Chat Panel (Figure 1-C) in Line Loop mode regarding artistic interpretation and other topics related to the selected line. The user can open a chat thread (ⓕ in Figure 1) for each line by clicking the 'Start Chat' button (See Figure 3, left). Once a chat thread is open, users can either select from four shortcut buttons (Meaning, Glossing, Emoting, Timing; ⓖ in Figure 1) or type their input directly into the chat box. In Global Mode, previous chat threads are displayed in a condensed format for fast navigation. To invite the user to the discussion on lines that are potentially challenging to translate, the system analyzes the lyrics in advance (*c.f.*, Section 4.5) and puts a light bulb icon (ⓑ in Figure 4) on the noteworthy lines (ⓐ in Figure 4). When the user initiates a chat thread for these lines, the chatbot directly starts a discussion on the meaning of the line (ⓒ in Figure 4).

4.3 Conversational Design: 4 Main Discussion Topics

The dialogue interaction is structured around key discussion topics designed to support users throughout the translation process. The process begins with a preprocessing stage, where ELMI analyzes the lyrics to identify abstract or idiomatic language and generates a default gloss along with additional relevant data, such as emotional tones and timing cues. When a user requests glossing, ELMI suggests translations based on the default gloss, contextualized to align with the intended meaning of the lyrics. If a user provides their own gloss, the ELMI recognizes this input, updating its suggestions to align with the user's input. This capability is managed through specialized chatbot modules that handle different discussion topics such as meaning extraction, gloss refinement, emoting, and timing:

- ❓ **Meaning** facilitates discussions about the message of the song, addressing poetic and ambiguous expressions, and considering variations due to different cultures, ages, and locations.



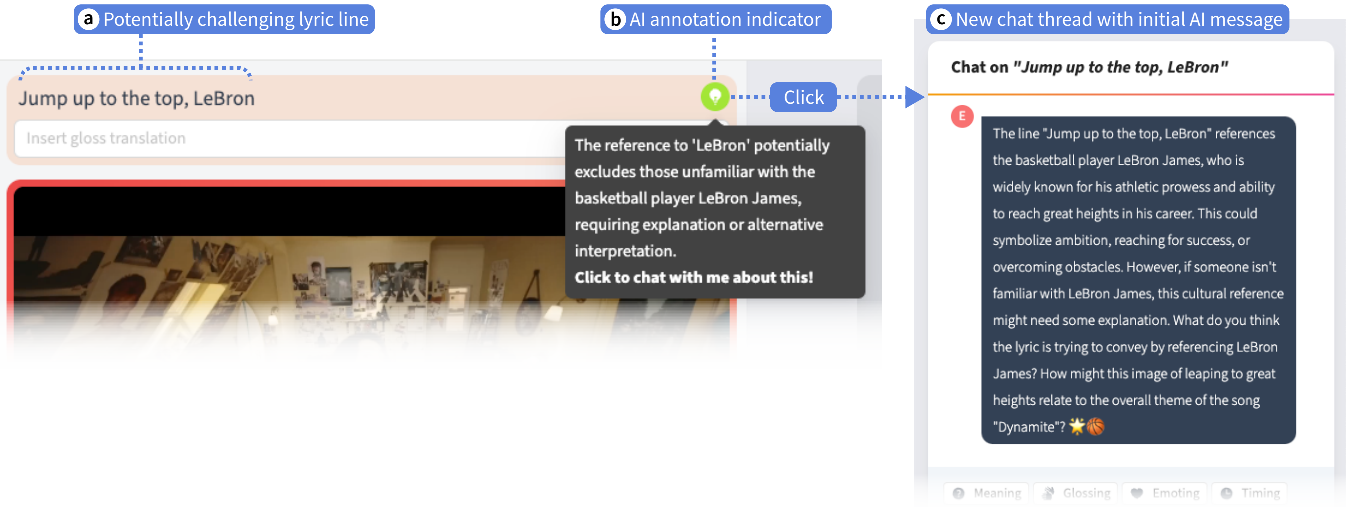
ⓐ Relative timeline in verse scope

ⓑ Play control for the current segment

ⓒ Audio-synced lyrics highlighting the current word

ⓓ AI suggestion of alternative translation

Figure 3: The lyric line translation pane. ELMI offers rich visual feedback to convey the song's timing; the user can check relative music position ⓐ at a verse level, as well as the line level ⓑ. While the music is being played, the corresponding lyric words highlight to enhance the user's sense of timing ⓒ. When the user is typing in the gloss, the system provides real-time suggestion of alternative translations in varied lengths ⓓ.


Jump up to the top, LeBron
Insert gloss translation
Click
Chat on "Jump up to the top, LeBron"
E
Meaning
Glossing
Emoting
Timing

ⓐ Potentially challenging lyric line

ⓑ AI annotation indicator

ⓒ New chat thread with initial AI message

The reference to 'LeBron' potentially excludes those unfamiliar with the basketball player LeBron James, requiring explanation or alternative interpretation.
Click to chat with me about this!

The line "Jump up to the top, LeBron" references the basketball player LeBron James, who is widely known for his athletic prowess and ability to reach great heights in his career. This could symbolize ambition, reaching for success, or overcoming obstacles. However, if someone isn't familiar with LeBron James, this cultural reference might need some explanation. What do you think the lyric is trying to convey by referencing LeBron James? How might this image of leaping to great heights relate to the overall theme of the song "Dynamite"? 🌟🏀

Figure 4: ELMI analyzes the lyrics in advance and marks noteworthy lines potentially challenging to translate ⓐ; In this case, an American basketball player, 'LeBron' may not be recognized by users unfamiliar with the US sports scene. When the user hovers over the annotation indicator ⓑ, it shows a tooltip that invites the user to the discussion. If the user starts a chat thread by clicking the indicator, the AI will start the discussion directly ⓒ.

- 👋 **Glossing** focuses on discussing sign selection, including non-manual signals (NMS). When there's no user input gloss, ELMI suggests ways to sign, while there's user input, ELMI suggests ways to improve signs.
- ♥ **Emoting** centers on discussing the tone, vibe, and emotion of the song, encouraging conversations to explore ideas by using facial expression and body language.
- 🕒 **Timing** addresses questions about adjusting the length of the gloss, offering users estimated timing without explicit numbers, multiple options for gloss length.

4.4 Interacting with ELMI

In this section, we illustrate how ELMI can be used through a scenario featuring Cindy, a Deaf song-signer with five years of experience. Cindy is preparing to interpret BTS's "Dynamite" in real-time at an upcoming concert and decides to use ELMI to create accurate glosses for the song.

After creating a new song project, Cindy enters the translation page. To familiarize herself with the song, she plays the music video, finding ELMI's feature of highlighting sung words in real-time particularly helpful for understanding the song's timing. Cindy begins her translation process by focusing on the first line of the song (ⓓ in Figure 1). As she clicks on the line, the *Information Panel* updates to provide a Mood and Performance guide (🅰 in Figure 1). She confirms that her feelings about the song align with ELMI's interpretation, which suggests a joyful and uplifting mood (ⓐ in Figure 1). ELMI also offers guidance on non-manual signals (NMS) such as facial expressions, advising Cindy to "`raise your eyebrows and smile widely to emphasize the feeling of being in the stars.`" (ⓑ in Figure 1). Reflecting on the lyrics "*'Cause I, I, I'm in the stars tonight*," Cindy writes her initial gloss as "BECAUSE I PARTY TONIGHT." (ⓓ in Figure 1.)

However, uncertain about the ❓ **meaning** of "stars," she opens a new chat thread in the *Chat Pane* and asks ELMI for clarification. ELMI suggests that "stars" could be a feeling of being extraordinary, celebrated, or having a bright moment (ⓕ in Figure 1). Taking this into account, Cindy revises her gloss to "BECAUSE I SHINE, ENJOY PARTY TONIGHT." She then synchronizes her gloss with the music, replaying the video in a loop. Noticing that her gloss is too long, Cindy asks ELMI for 🕒 **timing** advice and trims her gloss to "I SHINE, PARTY TONIGHT," confirming that it fits the song's timing.

When Cindy encounters the line "Jump up to the top, LeBron," she's unsure who "LeBron" refers to. She checks ELMI's tooltip (ⓑ in Figure 4), which annotates that "LeBron" refers to the basketball player LeBron James, noting that this cultural reference may require additional context or finger spelling. Finding this line challenging to translate into ASL, Cindy clicks the 👋 **Glossing** button to get a suggestion from ELMI. ELMI suggests "JUMP TOP F-S (FingerSpelling) 'L-E-B-R-O-N'," but Cindy opts to use a classifier to depict a basketball action, so she types "JUMP TOP CL-5 (basketball shooting)." ELMI's feedback confirms that her translation maintains the flow and meaning of the original line. Cindy then clicks ♥ **Emoting** and incorporates ELMI's suggestions to express the energy of jumping through her performance.

Cindy continues to gloss the song by line, interacting with ELMI throughout the process. Once she finalizes her gloss, she uses ELMI to play the music video from the beginning, signing along with the song and using her glosses as a guide.

4.5 LLM-driven Conversational Pipelines

The conversation system in ELMI leverages OpenAI's GPT-4 for preprocessing, intent classification, and chatbot interactions. The system uses tailored prompts specifically designed for each function to ensure accurate and contextually relevant responses. Refer to Appendix A.2 for an example of the prompt used for the discussion with the chatbot.

4.5.1 Lyric Analysis. When a user creates a new project, ELMI preprocesses the song by running the reference lyrics and metadata through four inference modules, as shown in the pipeline diagram (Fig. 5). The Line Inspector (Ⓑ in Figure 5) annotates each line for challenges like poetic, cultural, or mismatched meanings. Based on these notes (Ⓒ in Figure 5), the Base Gloss Generator (Ⓓ in Figure 5) creates default glosses for all lines. Using these glosses (Ⓔ in Figure 5), the Performance Guide Generator (Ⓕ) provides performance recommendations for gestures and expressions (Ⓖ in Figure 5), while the Alternative Gloss Generator (Ⓗ in Figure 5) produces alternative glosses and longer or shorter versions for all lines (Ⓘ in Figure 5). The Performance Guide Generator also generates a new performance guide for the line when the user updates its translation.

4.5.2 Intent classifiers. User input is processed either through shortcut buttons (ⓓ in Figure 1)) or direct input. When shortcut buttons are used, the system passes the intent directly. For open-ended queries, the intent classifier identifies the input and categorizes it into one of the four discussion topics.

4.5.3 Chatbot. All user queries are saved, allowing the chatbot to remember previous conversations and use the dialogue history for future interactions. Based on the classified intent, the chatbot selects prompts in two ways. (1) If there is no user-generated gloss, it uses base inference results from the pre-processing stage. (2) If a user-generated gloss exists, the prompts are adapted to incorporate the gloss as input for further refinement.

ELMI's prompt input is designed to facilitate thoughtful interaction. As an active listener, ELMI encourages users to think critically, posing reflective questions about the lyrics and glosses. ELMI's role is to assist users in improving their sign performance, while its personality remains defined by clear communication, humor, empathy, and sensitivity. ELMI keeps the conversation focused on lyric interpretation while gently redirecting off-topic discussions and offering continuous support. Responses are designed to be concise, engaging, and thought-provoking.

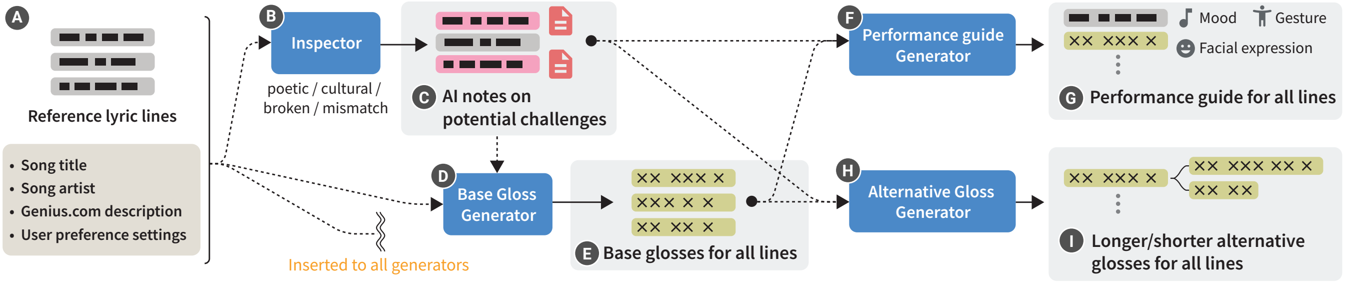

A
Reference lyric lines
Song title
Song artist
Genius.com description
User preference settings
B
Inspector
poetic / cultural / broken / mismatch
C
AI notes on potential challenges
D
Base Gloss Generator
Inserted to all generators
E
Base glosses for all lines
F
Performance guide Generator
Mood
Gesture
Facial expression
G
Performance guide for all lines
H
Alternative Gloss Generator
I
Longer/shorter alternative glosses for all lines

Figure 5: A pipeline for lyric analysis, which is part of pre-processing a song when the user creates a new project. Given the reference lyrics and metadata about the song and the user preference Ⓐ, the pipeline chains four LLM inference modules (Ⓑ, Ⓓ, Ⓕ, and Ⓖ) to generate notes on potential challenges when translating specific lines Ⓒ, base gloss translation Ⓔ, performance guides for lines Ⓖ, and the longer and shorter versions of each gloss line Ⓘ.

Table 1: Demographics, description, communication methods, and song-signing experience of our main study participants.

PID	Age/Gender	Description	Role	Song signing	Communication	Language
P1$^{d/D}$	37/M	Deaf (Profound/hearing aids)	Hobbyist (Church worship)	Always/4 years	Sign or talk (not at the same time)	ASL/fluent
P2$^{d/D}$	45/F	Deaf (Profound/hearing aids)	Deaf artist	Sometimes/25 years	Sign or talk (not at the same time)	ASL/native
P3$^{d/D}$	48/F	Deaf (Profound/hearing aids)	Official ASL interpreter	Often/30 years	Sign or talk (not at the same time)	ASL/native
P4$^{d/D}$	36/M	Hard of hearing (Profound/hearing aids)	Deaf artist	Often/14 years	Sign or talk (not at the same time)	ASL/fluent
P5$^{d/D}$	32/M	Hard of hearing (mild/none)	Official ASL interpreter	Often/15 years	Sign or talk (not at the same time)	ASL/fluent
P6$^{d/D}$	43/F	Hard of hearing (mild/none)	Hobbyist (Church worship)	Often/4 years	Talk only	PSE/moderate
P7^{Hg}	36/M	Hearing	Freelance ASL interpreter	Sometimes/4 years	Talk only	PSE/fluent
P8^{Hg}	27/F	Hearing	Hobbyist	Sometimes/12 years	Talk only	PSE/moderate
P9^{Hg}	25/M	Hearing	Deaf school teacher	Sometimes/4 years	Sign or talk (not at the same time	ASL/moderate
P10^{Hg}	36/Non-binary	Hearing	ASL artist	Always/12 years	Sign and talk at the same time	ASL/fluent
P11^{Hg}	67/F	Hearing	Hobbyist (Church worship)	Often/20 years	Sign or talk (not at the same time)	ASL/moderate
P12^{Hg}	36/M	Hearing	Hobbyist	Rarely/7 years	Talk only	ASL/novice
P13^{Hg}	30/F	Hearing	Official ASL interpreter	Rarely/8 years	Sign or talk (not at the same time)	ASL/fluent

4.6 Implementation

We developed ELMI's interface as a web application using TypeScript with React.js. The backend is powered by Python and built with the FastAPI framework, which serves as a REST API to manage both data and computational pipelines. These pipelines were implemented using LangChain to streamline multiple steps of LLM inferences, with data validation handled by Pydantic. A SQLite database is used to track the status of translations, messaging, and analysis results.

To provide fine-grained, real-time visual feedback for lyrics while playing audio, we used word-level timestamps (*i.e.*, start and duration) of lyrics automatically generated through sequence matching and automatic speech recognition (ASR). The pipeline leverages reference lyrics from Genius.com [39] lyric database and obtains word-level timestamps matching the lyrics with the ASR transcription results from OpenAI Whisper API [55], which provides word-level timestamps for transcription. To improve accuracy, we first segmented the audio by lines using the subtitle information in the song's official YouTube video and generated word-level timestamps for each line. (Refer to Appendix A.3 where we describe the detailed pipeline steps.)

5 User study

We conducted an exploratory user study with ELMI, employing observation methods to examine how a Large Language Model (LLM) chat agent can support song-signing gloss creation. To ensure the effectiveness of the system and the study protocol, we refined both through two pilot sessions with hard-of-hearing song-signers. The study protocol was approved by our Institutional Review Board.

5.1 Participants

We recruited *13* song-signers (P1–13) through various channels, including social media platforms such as Facebook groups and Reddit, email lists, word of mouth, and snowball sampling. Three participants self-identified as deaf, three as hard of hearing, and the remaining seven as hearing. Our inclusion criteria required participants to (1) have completed at least three song-signing projects before the study, (2) be a minimum of 18 years old, (3) be comfortable reading and writing in English, and (4) use American Sign Language (ASL) or Pidgin Signed English (PSE) in their song-signing work. Table 1 summarizes the demographic of our study participants. Participants were aged between 25 and 67 years ($M = 38.3$) and included 9 males, 4 females, and 1 non-binary individual. Ten out of 13 participants communicated primarily in ASL, with others using Pidgin Signed English (PSE). On average, participants had 12.2 years of experience in song-signing, with their involvement ranging from 4 to 30 years, driven by professional and personal motivations, ranging from ASL interpreters and artists to hobbyists and Deaf school teachers.

5.2 Songs for Translation

Our study involved translating two songs. First, participants translated the same song (referred to as the **Song**^{Assigned} hereinafter) so that we could observe common interaction patterns and diversity of translation. Second, each participant chose one song (referred to as the **Song**^{Selected} hereinafter) they wanted to translate.

For the Song^{Assigned}, we selected "Butter" by "BTS" [5], considering its moderate complexity, clear emotional tone, and being free

---

[5] https://www.youtube.com/watch?v=WMweEpGlu_U

of explicit content. The song conveys themes of self-discovery, resilience, and love [62, 75]. The song's lyrics, which include multiple types of figurative language such as metaphors, similes, hyperbole, and personification [5, 22, 26], which was suitable to test ELMI's ability to interpret and gloss complex expressions.

For the Song^{Selected}, we asked participants to select up to three candidate songs that meet the following criteria: (1) written in English, (2) with an official music video available on YouTube, and (3) with reference lyrics available online. Participants were encouraged to choose preferred songs, as this would likely lead to meaningful interactions. The list of selected songs and YouTube videos used in the study is provided in Appendix A.4.

5.3 Study Procedure

5.3.1 Pre-study Preparation. Before the main study session, participants completed a pre-study survey, which collected demographic information, hearing description, song-signing experience, ASL expertise, and details about the tools and methods they used for song-signing and gloss creation. Participants also submitted three candidates for the Song^{Selected}.

5.3.2 Main Study Session. Each participant engaged in a 2-hour main study session remotely on Zoom. Depending on their hearing capability, the session was carried on via spoken language with live captioning, Zoom chats, or with an ASL interpreter. The Zoom session was audio- and video-recorded for analysis.

Briefing and Tutorial. We first described the goal of our study and the study protocol, covering the overall process of creating a project, glossing, and chatting. As a tutorial for the ELMI interface, participants shared their screen and tried ELMI with the example song, "Viva La Vida" by Coldplay. Participants were allowed to practice with the tool until they felt confident about the interface. This phase took about 20 minutes.

Song Translation. Participants created glosses using ELMI for two projects, translating Song^{Assigned} and Song^{Selected}. Due to time constraints, we asked participants to translate only the first verse of each song. We allowed up to 30 minutes for each song but also asked not to rush to finish within the time limit and to prioritize the translation quality as usual.

Debriefing. We conducted a semi-structured interview around 30 minutes at the end of the session. We asked participants about their experience with ELMI, the quality of translation they created, the benefits and concerns of ELMI, and follow-up questions based on our observations in the translation phase. We also asked them how they envision ELMI to be applied to their workflow and their behavioral intention to use the tool after the study. We offered participants a 75 USD Amazon gift card for their participation.

5.3.3 Post-study Assignments. Due to the time constraints, most participants did not finish the two songs during the session. Hence, we asked participants to finish verse 1 of the Song^{Assigned} after the study session, offering additional incentives of 25 USD. Participants also optionally completed the Song^{Selected} and submitted a post-study survey, where we asked about their experience of using ELMI in their own environment. This survey also included a 5-point Likert scale questionnaire consisting of questions derived from the Technology Acceptance Model (TAM) [18] regarding the self-assessed quality of translation, ease of use, and willingness to use ELMI. Participants who finished this optional task were compensated with an additional 80 USD. All participants except P5$^{d/D}$ and P13^{Hg} completed post-study tasks.

5.4 Analysis

Glossing. We analyzed the translation results of the Song^{Assigned} from 11 participants who completed the first verse. Key metrics such as the number of lines completed and the average word count were analyzed using Pandas[6] in Python.

Chats. We examined chat threads of both Song^{Assigned} and Song^{Selected} from 11 participants who completed post-study assignments. We defined a *turn* as a single message exchange, with *user turns* representing participant messages and *system turns* representing those by ELMI. Using the Pandas Python package, we computed various descriptive statistics, such as the total number of turns per session. Additionally, we categorized conversation turns using a top-down approach based on four discussion topics, and researchers independently coded the dialogues of three participants.

Interviews and Surveys. We also analyzed debriefing transcripts, post-study surveys, and video recordings and screen captures, to understand participants' perspectives and expectations regarding ELMI. Thematic analysis [9] was employed to code and examine the transcripts. Initial coding was performed by two researchers independently, followed by iterative discussions to resolve discrepancies and refine the codes. Recurring codes were then grouped into broader categories and further distilled into key themes through team discussions. Three key themes emerged: interactions with ELMI, including the four discussion topics; perceptions of ELMI's tone and its influence on translation ownership; and the benefits and challenges experienced while using ELMI.

6 Results

This section presents the study's findings across five key areas: (1) Translation Outputs, (2) Conversations with ELMI, (3) Perceptions of ELMI and (4) Strengths and Limitations of Using ELMI.

6.1 Translation Outcome

We present the glossing results for Song^{Assigned}, BTS's "Butter" to examine how ELMI affected the diversity of translations produced by participants for the same lyrics. We analyzed the first three parts (Verse 1, Pre-chorus, Chorus), comprising 19 lines with a total of 105 words. Among the participants, six (3 d/Deaf, 3 hearing) used ASL, while five (2 d/Deaf, 3 hearing) used PSE. Participants spent varied time translating lyrics during the main session, completing between 3 and 19 lines, with an average completion of 9.82 lines (SD = 5.25, min = 2, max = 19).

Table 2 summarizes the translation result of the first verse of Song^{Assigned} by 11 participants who completed the post-study assignments. Participants produced a wide variety of gloss translations, incorporating non-manual signals (NMS), such as classifiers, visual vernacular, facial expressions, and body language. On average, participants used 3.72 signs per line (SD = 1.38, min = 1.9, max = 5.72), with 0.44 NMS per line (SD = 0.91, min = 0.09,

---

[6] https://pandas.pydata.org/

max = 1.09). There was a high individual variance in both the number of signs and NMS. We did not observe a significant difference between the d/Deaf and hearing participants (3.61 signs and 0.75 NMS per line for d/Deaf, and 3.81 signs and 0.18 NMS per line for hearing participants), although d/Deaf participants tended to use NMS more actively.

To assess the diversity of translation among participants, we calculated the average overlap coefficient[7] of pairwise combinations of manual signs for each lyric line (See Table 2, 'Average Overlap'). The average overlap coefficient of ASL signs of 19 lines was 36.36% (SD = 20.53%, min = 0%, max = 72.22%), and that of PSE was 38.39% (SD = 21.62%, min = 0%, max = 80.38%), suggesting that participants shared around one-third of signs for the same line. For instance, when translating line 2, "*Like a criminal undercover*," PSE participants seemed to assign literal signs such as "LIKE SECRET THIEF" [P7^{Hg}]. In contrast, ASL participants produced more diverse and distinct signs such as "ITSELF THEFT OVERLOOK" [P1$^{d/D}$] or "PERSON DANGEROUS DISGUISE" [P3$^{d/D}$].

6.2 Conversation with ELMI

Overall, participants opened a total of 222 threads, with 114 threads from Song^{Assigned} and 108 from Song^{Selected}. These threads comprised 1,346 messages—730 from Song^{Assigned} and 616 from Song^{Selected}, featuring 563 user messages (308 from Song^{Assigned}, 255 from Song^{Selected}) and 775 system messages (414 from Song^{Assigned}, 361 from Song^{Selected}). For Song^{Assigned}, participants opened an average of 10.36 threads from 19 lines (SD = 5.32, min = 3, max = 17), with each thread containing 5.91 messages (SD = 1.97, min = 3.5, max = 10.53). As for Song^{Selected}, participants opened an average of 9.82 threads from 47.36 lines (SD = 7.07, min = 1, max = 23), with each thread containing 5.14 messages (SD = 1.61, min = 3, max = 7.73) (See Table 3). To interact with ELMI, participants either clicked on one of four shortcut buttons (See ⓓ in Figure 1) or typed manual inputs. Participants used both shortcut message and manual message in a similar manner (See Table 3). Regarding the four discussion topics, participants rated them as both useful (See Figure 6) and frequently used (See Table 4) in the following order: 👋 glossing, ❓ meaning, ♥ emoting, and 🕒 timing.

❓ **Meaning** was frequently discussed at the beginning of their threads to explore how to interpret specific words, 7.9 times (SD = 6.25, min = 0, max = 17) for Song^{Assigned} and 4.55 times (SD = 4.1, min = 0, max = 12 for Song^{Selected} (See Table 4). ❓ Meaning was rated 3.81 on average (SD = 1.25, min = 2, max = 5), where d/Deaf participants gave 4 (SD = 1.22), and hearing participants gave 3.66 (SD = 1.37) (See Figure 6). P6$^{d/D}$ engaged with ELMI to interpret "Cool shade stunner". ELMI broke down the metaphor and encouraged P6$^{d/D}$ to consider the symbolic meaning and how it could be visually represented (See Dialogue 1).

Participants valued its ability to clarify and contextualize song lyrics, enabling them move beyond literal translations to more nuanced, expressive interpretations. As P2$^{d/D}$ noted, "*In the past, my translations were usually very literal. I didn't think about the abstract meaning. But now, with ELMI, I realize I can create more*

Table 2: Summary of participants' translation of the first verse of Song^{Assigned}, including the word count of gloss lines and the average of pairwise overlap coefficient across participants' gloss words for the same line. 0% of average overlap means that all glosses from participants include distinct words.

Line	Lyric	Lang	Example gloss	Gloss word count				Avg. overlap
				Min	Max	Mean	SD	
1	Smooth like butter	ASL	ME SAME-AS BUTTER SMOOTH (P10^{Hg})	2	5	3.50	1.22	67.33%
		PSE	SMOOTH LIKE BUTTER (P6$^{d/D}$, P8^{Hg}, P11^{Hg})	3	5	3.80	1.10	62.00%
2	Like a criminal undercover	ASL	PERSON DANGEROUS DISGUISE (P3$^{d/D}$)	2	3	2.83	0.41	0.00%
		PSE	LIKE SECRET THIEF (P7^{Hg})	3	5	3.60	0.89	32.50%
3	Gon' pop like trouble	ASL	GUN [LCL"shoot"] (P2$^{d/D}$)	1	4	2.83	1.17	11.11%
		PSE	APPEAR LIKE TROUBLE (P8^{Hg}, P11^{Hg})	3	4	3.20	0.45	50.00%
4	Breakin' into your heart like that	ASL	PENETRATE HEART HEART [CL "pump"] (P10^{Hg})	2	5	3.17	1.17	53.33%
		PSE	ME ENTER YOUR HEART (P6$^{d/D}$)	3	5	4.40	0.89	22.33%
5	Cool shade stunner	ASL	[CL:1 "person wearing sunglasses, looking cool"] [HEAD-tilt "same as music video"] (P2$^{d/D}$)	0	4	2.83	1.17	22.22%
		PSE	COOL SHADE AWESOME PERSON (P7^{Hg})	1	4	3.00	1.22	9.17%
6	Yeah, I owe it all to my mother	ASL	[HEAD-nod] THANK MOTHER (P2$^{d/D}$)	2	6	3.50	1.38	21.67%
		PSE	ME OWE MOM EVERYTHING (P8^{Hg})	3	7	5.40	1.82	27.86%
7	Hot like summer	ASL	SUNSHINE [CL:5 "coming from me"] (P10^{Hg})	1	4	2.67	1.03	55.56%
		PSE	HOT LIKE SUMMER (P6$^{d/D}$, P8^{Hg})	3	5	4.00	1.00	79.83%
8	Yeah, I'm makin' you sweat like that	ASL	YOU SWEAT YES (P12^{Hg})	2	5	3.67	1.21	46.11%
		PSE	ME CAUSE YOU SWEAT (P4$^{d/D}$, P6$^{d/D}$, P8^{Hg})	2	8	4.60	2.79	40.36%
9	Break it down	ASL	RELAX ENJOY (P2$^{d/D}$)	1	4	2.00	1.10	6.67%
		PSE	DANCE (P8^{Hg})	1	3	1.80	0.84	0.00%

---

[7] Calculated with the length of intersection of the word sets A and B divided by the length of min(A, B)

Table 3: Average number of messages per thread and the breakdowns of user-generated messages from the shortcut buttons (shortcut messages) and those typed manually (manual messages) per participant.

	$\text{Song}^{Assigned}$		$\text{Song}^{Selected}$	
# of messages per thread	5.91	(SD = 1.97)	5.14	(SD = 1.61)
d/Deaf only	5.75	(SD = 1.72)	4.74	(SD = 1.98)
hearing only	6.03	(SD = 2.31)	5.47	(SD = 1.34)
# of shortcut messages	13.36	(SD = 11.93)	11.55	(SD = 9.77)
d/Deaf only	7.20	(SD = 8.50)	5.40	(SD = 7.09)
hearing only	18.5	(SD = 12.55)	16.67	(SD = 9.03)
# of manual messages	14.64	(SD = 17.98)	11.64	(SD = 12.71)
d/Deaf only	16.2	(SD = 19.05)	8.40	(SD = 12.58)
hearing only	13.33	(SD = 18.76)	14.33	(SD = 13.32)

Table 4: Average number of messages per participant by discussion topic.

	$\text{Song}^{Assigned}$		$\text{Song}^{Selected}$	
# of Meaning messages	7.90	(SD = 6.25)	4.55	(SD = 4.10)
d/Deaf only	9.20	(SD = 6.87)	5.00	(SD = 5.00)
hearing only	6.83	(SD = 6.11)	4.17	(SD = 3.66)
# of Glossing messages	14.64	(SD = 13.98)	11.90	(SD = 11.28)
d/Deaf only	10.00	(SD = 7.78)	5.00	(SD = 5.00)
hearing only	18.50	(SD = 6.11)	17.67	(SD = 12.13)
# of Emoting messages	4.10	(SD = 4.57)	4.63	(SD = 8.20)
d/Deaf only	3.00	(SD = 3.08)	1.80	(SD = 2.49)
hearing only	5.00	(SD = 5.66)	7.00	(SD = 10.71)
# of Timing messages	1.36	(SD = 1.12)	2.10	(SD = 2.07)
d/Deaf only	1.20	(SD = 1.30)	2.00	(SD = 2.55)
hearing only	1.50	(SD = 1.05)	2.17	(SD = 1.35)

ELMI	*The phrase "Cool shade stunner" is poetic, employing a metaphor to convey a style or attitude that may be difficult to interpret directly into sign language.* *What do you think the author means by "cool shade"?*☘ *Is it about being physically cool or could it symbolize something else, like a relaxed, confident style?*✌ *Is there a "cool shade" sign that might capture that feeling?*
$\text{P6}^{d/D}$	*I think it's symbolic of someone being really suave and smooth to get into someone's heart.*
ELMI	*That's an interesting perspective, [P6's nickname]!* ❣ *The idea of someone being suave and smooth certainly fits the essence of the lyric. Now, thinking about translating that into PSE, how would you visually express "suave" and "smooth"?*✨ *Do you think you'd use certain movements or expressions to convey that relaxed confidence we talked about?*

Dialogue 1. Example conversation on ❓ Meaning

visual interpretations. It's not just a word-for-word translation but something more engaging." The discussion was particularly effective in interpreting poetic language, offering alternative perspectives that helped participants explore different interpretations.

👋 **Glossing** was the most frequently used and highly rated topic: 14.64 times (SD = 13.98, min = 1, max = 50) for $\text{Song}^{Assigned}$ and 11.9 times (SD = 11.28, min = 0, max = 38) for $\text{Song}^{Selected}$ (See Table 4). It had an average rating 3.9 (SD = 1.04, min = 2, max = 5), 4 from d/Deaf participants (SD = 1), and 3.83 from hearing participants (SD = 1.17) (See Figure 6). Song-signers often relied on it to get feedback on their glosses. For example, P7^{Hg} asked ELMI for guidance on a specific signing selection, and ELMI suggested several ways to combine signs to enhance expressive quality (See Dialogue 2).

P7^{Hg}	*Using the "chilled girl" sign along with the "butter" sign is a good match?*
ELMI	*Using "CHILLED GIRL" with the "BUTTER" sign could indeed be a creative match, as it blends the relaxed and smooth concepts together in a visually expressive way.* *Here's how you might combine them:* *1.**"CHILLED"** - Sign "RELAX" or "COOL" to capture the essence of being chilled.* *2. **"GIRL"** - Sign "GIRL" with a thumb brushing the cheek.* *3. **"BUTTER"** - Add the sign for "BUTTER," which you can do by mimicking a smooth, spreading motion with your hand, like spreading butter on bread.* *So your combined gloss might look like this: **"RELAX GIRL BUTTER"** or **"COOL GIRL BUTTER"** with smooth, flowing motions to emphasize the cool and effortless vibe.* ❄ ♨ *Does this combination feel like it captures the meaning and essence you're going for?* *Would you like to explore any additional elements to enhance it?*

Dialogue 2. Example conversation on 👋 Glossing.

Both d/Deaf and hearing participants found the 👋 Glossing valuable for exploring alternative options for specific phrases or signs, enabling creativity and adjustments based on the context and tone of the song. The feedback provided by ELMI was also well-received, as P9^{Hg} noted, *"Glossing is more effective when I suggested my own ideas, as ELMI explains synonymous words and also offers suggestions on how to convey the message."*

♥ **Emoting** was used 4.1 times (SD = 4.57, min = 0, max = 14) for $\text{Song}^{Assigned}$ and 4.63 times (SD = 8.2, min = 0, max = 28) for

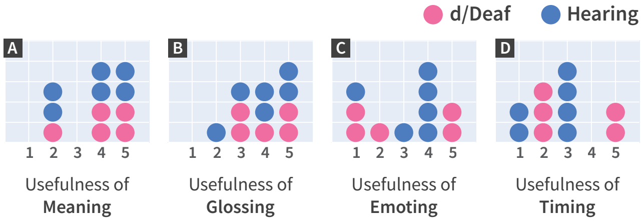


Figure 6: Likert-scale ratings of the 4 discussion topics (1: not useful 5: extremely useful).

Song^{Selected} (See Table 4). ♥ Emoting had an average rating 3.09 (SD = 1.58, min = 1, max = 5), where d/Deaf participants gave 2.8 (SD = 2.04), and hearing participants gave 3.33 (SD = 1.21) (See Figure 6). It provided participants with suggestions on how to enhance the performance of specific lyrics through facial expressions and body language. For example, P3$^{d/D}$ asked about using "raised eyebrows" after finalizing her gloss "SUDDENLY DANGER" to translate the phrase "Gon' pop like trouble." (See Dialogue 3.)

P3$^{d/D}$	*RAISED EYEBROWS*
ELMI	*Raised eyebrows can definitely convey surprise and alertness, which fit perfectly with "SUDDENLY DANGER."* ❗ *Would you also consider adding a quick, sharp movement with your hands to emulate the suddenness?* *Or perhaps a slight lean forward to convey the immediate attention the situation demands?* *What do you think would pair well with those raised eyebrows to fully capture the impact?* 😊
P3$^{d/D}$	*YES*
ELMI	*Fantastic! Raised eyebrows paired with a quick, sharp hand movement and a slight lean forward should effectively convey "SUDDENLY DANGER."* ⚠️

Dialogue 3. Example conversation on ♥ Emoting.

Most hearing participants (N = 6) found the ♥ Emoting conversation valuable for its guidance on facial expressions and body language, helping them to convey emotions effectively. This feature was likened to "*Surrogacy in ASL* (P10^{Hg})" and "*Acting notes from a director* (P10^{Hg})", with an appreciation for its conceptual approach. However, most d/Deaf participants (N = 4), who already internalized sign language's emotional aspects, found the feature less relevant or redundant. P1$^{d/D}$ noted, "*Because I grew up as a Deaf person, I'm already expressive with my face.*"

🕒 **Timing** was used 1.36 times (SD = 1.12, min = 0, max = 3) for Song^{Assigned} and 2.1 times (SD = 2.07, min = 0, max = 6) for Song^{Selected} (See Table 4). 🕒 Timing was rated 2.73 on average (SD = 1.34, min = 1, max = 5), 3.2 by d/Deaf participants (SD = 1.64), 2.33 by hearing participants (SD = 1.03) (See Figure 6). Participants checked the timing of the glosses to align them with the rhythm and tempo of the song (See Dialogue 4).

P12^{Hg}	*Yes that sounds great! How about the timing?*
ELMI	*Great question about timing! Because ASL is so visual, you can match the gloss with the rhythm of the song seamlessly.* *Imagine signing "MELT" with a smooth, fluid motion, taking your time to show the melting process, then transitioning into "SMOOTH" just as effortlessly.* ☕ *You can slow down your signs to match the song's tempo.* *Does that match how you envision it? How do you think your audience will perceive the timing and the flow?*

Dialogue 4. Example conversation on 🕒 Timing.

Even though both d/Deaf and hearing participants did not discuss 🕒 Timing extensively, its explanatory value was appreciated. Some participants (N = 5) found 🕒 Timing particularly useful as a "*cross-check tool* (P7^{Hg})", especially when signing along with "*fast songs* (P6$^{d/D}$)", to ensure they were on the right track with the rhythm. The perceived importance of 🕒 Timing varied among users. Those with more experience or a natural sense of rhythm often found it less critical, while beginners or those less familiar with musical timing considered it more valuable. As P11^{Hg} noted, "*[...] it[Timing] is not as crucial as it is for hearing people who are more attuned to musical timing.*"

6.3 Perceived Attitude and Ownership

Reflecting on the conversation with the ELMI chatbot, participants reported various attitudes and stances of the chatbot. Participants generally found ELMI to be **encouraging and supportive** in the translation process, helping users feel more confident in their work, while it sometimes provided **critical and constructive** suggestions helping participants fine-tune their translations. P7^{Hg} and P4$^{d/D}$ noted that ELMI supported balancing between their initial ideas and refined ones: "*The suggestions helped me align my approach, providing a healthy middle ground if I was overthinking or missing something.* (P4$^{d/D}$)" ELMI was also **informative and guiding** by providing context that deepened participants' understanding of the songs. P7^{Hg} and P10^{Hg} highlighted how ELMI helped them grasp the essence of the songs, which enhanced their interpretations: "*There are times when I'm not aware of the context, but ELMI already understands it, guiding me setting the mood, facial expressions, and the story I should convey. Those suggestions are critical.* (P7^{Hg})"

In the post-study survey, participants highly rated their control or authorship over the final gloss, rating 4.63 out of 5 (SD = 0.5, min = 4, max = 5; see Figure 7-G). During debriefing, the majority of participants (N = 11) echoed this sentiment, noting that they retained control over their work. They appreciated that ELMI did not impose decisions but instead supported their autonomy throughout the translation process. Participants generally valued their own intuition and experience over ELMI's suggestions, particularly when they disagreed with the tool's input. They used ELMI's input as a starting point or an additional perspective rather than as a final directive. Consistently, participants expressed a strong sense of ownership over their glosses. P12^{Hg} noted, "*I would definitely give you credit, but it still feels like the work is mostly mine.*"

However, a few participants felt that ELMI sometimes overshadowed their expertise, leading to moments of self-doubt and frustration. P3$^{d/D}$ mentioned feeling discouraged when the AI suggested solutions they had already considered, creating a sense of competition rather than collaboration. Similarly, P11^{Hg} noted that ELMI's suggestions occasionally did not align with their intended interpretation, causing them to question their own decisions.

6.4 Strength and Drawbacks of ELMI

In the post-study survey, participants highly evaluated their satisfaction with the quality of translation they produced using ELMI, rating an average of 4.27 out of 5 on a Likert scale (SD = 0.65, min = 3, max = 5). Participants found ELMI easy to use, giving it an average rating of 4.36 out of 5 (SD = 0.5, min = 4, max = 5). d/Deaf participants rated it 4.2 (SD = 0.45), while hearing participants rated it slightly higher at 4.5 (SD = 0.55) (See Figure 7-C).

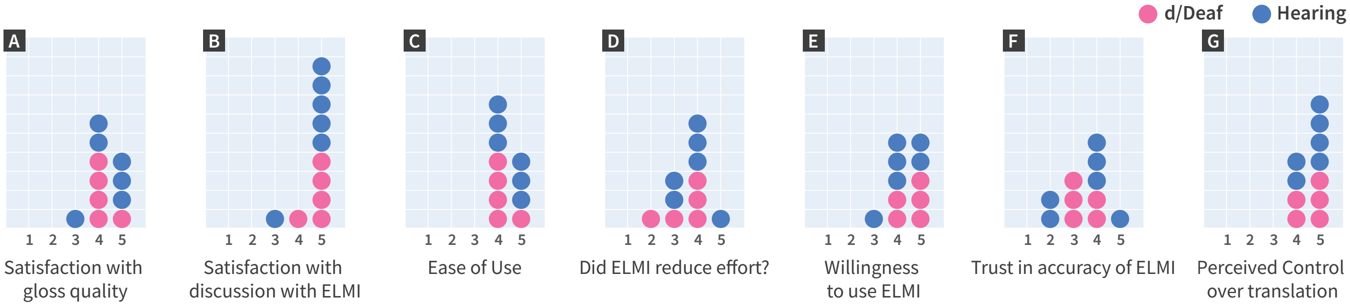


Figure 7: Likert-scale ratings results for quality of gloss, discussion, ease of use, reduction of effort, intention of use, accuracy, and control (1: negative, 5: positive)

In terms of reducing participants' effort, ELMI received an average rating of 3.63 ($SD = 0.8$, $min = 2$, $max = 5$), with d/Deaf participants giving it a 3.4 ($SD = 0.89$) and hearing participants rating it 3.83 ($SD = 0.75$) (See Figure 7-D).

However, participants were more critical of ELMI's accuracy, with an average rating of 3.45 out of 5 ($SD = 0.93$, $min = 2$, $max = 5$). d/Deaf participants rated the accuracy at 3.4 ($SD = 0.54$), while hearing participants rated it 3.5 ($SD = 1.22$) (See Figure 7-F).

6.4.1 How ELMI Supported the Translation Workflow. Participants expressed a strong interest in integrating ELMI into their workflow, with an average willingness rating of 4.36 out of 5 on the Likert scale ($SD = 0.67$, $min = 3$, $max = 5$). d/Deaf participants rated their willingness slightly higher, at 4.6 ($SD = 0.55$), while hearing participants rated it 4.17 ($SD = 0.75$). Figure 8 summarizes the main features of ELMI that participants mentioned as the top three favorites in the post-study surveys. In the following, we cover the noteworthy features.

Line-by-line Focus. Most of the participants ($N = 8$, 4 d/Deaf and 4 hearing) liked ELMI's ability to break down song lyrics into "lines", making the translation process more precise (See Figure 8). This was particularly useful for complex lyrics, where understanding the meaning is crucial. Additionally, looping specific lines with the video was received as useful for practice and refinement. P11^{Hg} mentioned ELMI's "*line-by-line setup made it easy to see how the person in the video is conveying emotions and actions, which is incredibly helpful for aligning translation with the intended expression.*"

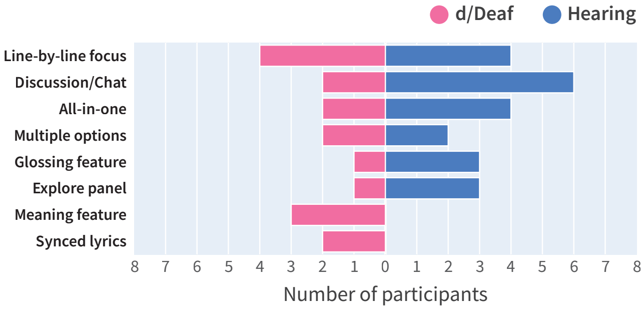


Figure 8: Distribution of participant responses (39 total) regarding the strengths of ELMI across three choices.

Interactive Discussion and Chat. As P9^{Hg} described, "*It doesn't feel like you're alone in interpreting*", discussing with ELMI was highly rated by participants ($N = 8$, 2 d/Deaf and 6 hearing) for providing a sense of independence and efficiency during the translation process, reducing their reliance on other people for feedback (See Figure 8). P2$^{d/D}$ remarked, "*[...] Without ELMI, I'd have to ask my interpreter how the voice sounds, which can be a long process. It really depends on the person I'm working with, and I have to trust them more. But ELMI gives me more confidence and independence.*" Although P13^{Hg} also noted that it may not entirely replace the need for direct consultation when facing difficult decisions, P10^{Hg} mentioned that ELMI serves as an effective "intermediary tool," helping to refine initial thoughts before seeking further input from others making the process more thorough and reflective.

ALL-IN-ONE workplace. The integration of lyrics, glosses, videos, and contextual information into a single platform was widely seen as a major advantage ($N = 6$, 2 d/Deaf and 4 hearing), streamlining the workflow and allowing participants to focus more on the translation itself (See Figure 8). P4$^{d/D}$ highlighted the convenience of this approach: "*Normally, I would have to navigate between different resources, which can be cumbersome. If I needed to verify something, I had to keep going back and forth.*"

Synced Lyrics with Music Video. Deaf participants particularly valued ELMI's ability to synchronize lyrics with the music, making the process of matching the song's tempo more accessible. Usually, this process involved filming themselves, using metronomes, and consulting with interpreters to ensure their sync was accurate. However, ELMI simplified this for deaf participants. P2$^{d/D}$ highlighted the effectiveness of the visual cues: "*I like the highlighted words in red that pop up. With the cues, I can easily follow the beat and see which word is currently being emphasized. As a deaf person, it's very helpful to see those and understand the song's tempo.*" P1$^{d/D}$ referred to this feature as "Karaoke methods": "*It allows me to see where the song is sung. I'm profoundly deaf and wanted to know this!*"

6.4.2 Challenges and Concerns of ELMI. Besides the benefits, participants also encountered challenges and shared various concerns about ELMI regarding the d/Deaf community and culture.

Loss of Human Interaction and Cultural Appropriation. Some participants (P3$^{d/D}$, P7^{Hg}) expressed concern that ELMI diminishes the "personal touch of human interaction (P3$^{d/D}$)". They missed the immediate feedback and personalized discussions that typically come from working with other people. Additionally, d/Deaf participants voiced concern that hearing individuals using ELMI might unintentionally engage in cultural appropriation, urging for careful usage: "*I think interpreting songs really needs to spotlight the Deaf community. If a hearing person uses technology like ELMI, it might lead to cultural appropriation. They might use it for their own benefit, without the intention of spreading sign language or supporting the Deaf community.*" (P2$^{d/D}$)

Participants suggested ways to ensure accuracy and help maintain the human element in interpreting songs. d/Deaf participants (P2$^{d/D}$, P3$^{d/D}$) suggested collaboration features to share their work with interpreters, take notes, and get feedback from others to refine their signs. Additionally, P10^{Hg} emphasized the importance of adding a disclaimer: *"This tool doesn't replace the expertise of a Deaf person. When signing music, it's important to consult with a Deaf person, even when using this tool."*, reminding users to consult with Deaf individuals during the translation process. Furthermore, P11^{Hg} highlighted the need to integrate feedback from the Deaf community regarding the use of AI in conveying nuanced expressions. "*The effectiveness of AI in conveying expressions is still uncertain. I believe a real person could convey these nuances better than AI. It might be useful to survey the Deaf community to understand their preferences whether they'd prefer real human demonstrations or if they'd be comfortable with AI-generated visuals.*"

English-centric Generation hampering Accuracy and Depth of Translation. ELMI was seen to be "English-focused (P3$^{d/D}$)", with its concepts presented from a "non-Deaf perspective (P10^{Hg})". Participants (P3$^{d/D}$, P10^{Hg}, P13^{Hg}) noted that ELMI might be better suited for Pidgin Sign Language (PSE) or Signed Exact English (SEE). P13^{Hg} noticed ELMI often adhered English word order, lacking deeper integration with ASL grammar and nuances: "*I'm noticing that ELMI's suggestions for the gloss are still following English word order, even though it was set to ASL. I understand that it's challenging for AI to grasp ASL grammar. For example, using signs like 'stealing person' and the concept of 'undercover' seems more like PSE than ASL.*"

Additionally, ELMI's responses were sometimes seen as too shallow, particularly when dealing with poetic expressions. Participants (P9^{Hg}, P11^{Hg}, P13^{Hg}) felt that ELMI often failed to fully grasp the deeper meanings behind certain phrases. P11^{Hg} concern that ELMI might not fully capture the nuances, "*I'm unsure if ELMI can handle such specific phrases without seeing them in action.*", given that ELMI is not actually observing their performance. P12^{Hg} also questioned the accuracy of ASL suggestions that ELMI might not be drawing from sign dictionaries: "*It may be suggesting valid glossing, but also could be hallucinating.*" This uncertainty led to a lack of trust in ELMI's recommendations.

To make ELMI more culturally and linguistically appropriate, participants offered several key suggestions. Many participants ($N = 7$) recommended integrating an ASL dictionary that includes image or video demonstrations, as well as detailed hand shapes, which would be especially beneficial for novice song-signers. For instance, P3$^{d/D}$ remarked, "*Seeing examples of signs could bridge the gap for students who are still learning ASL. For example, if I could click on something and see the sign for a word like 'butter' or 'smooth,' it would be very helpful.*"

Lose of Context due to Line-by-line Focus. Although the concept of line-by-line focused translation was generally well-received, some participants raised concerns on it; they felt focusing on a line basis could overlook the broader context of the song and disrupt the continuity of ideas. As P5$^{d/D}$ mentioned: "*If someone were using ELMI, they might focus too much on translating line by line, instead of considering the overall structure.*" As such, P4$^{d/D}$ and P5$^{d/D}$ suggested supporting multiple-line selection, which allows users to consider the broader context of the song, addressing the narrow focus of the line-by-line approach. P4$^{d/D}$ explained that for slower songs, the current single-line view works well. But for faster genres like hip hop, viewing two lines at once would help the user better understand connections between lines, such as punchlines and their surrounding context.

Insufficient Critical Feedback. Four participants pointed out that ELMI was sometimes seen as too supportive, often providing redundant feedback that did not challenge the participants' interpretations. This led to concerns about the tool not being critical enough, which could hinder improvement. Participants (P3$^{d/D}$, P5$^{d/D}$) also worried that less experienced signers might misuse ELMI's suggestions, potentially adopting incorrect interpretations. As P3$^{d/D}$ noted, "*This could serve as a shortcut, even though [they] haven't fully studied.*"

7 Discussion

In this section, we discuss lessons learned from the design and evaluation of ELMI. We also reflect on the implications for better supporting artistic sign language translation in an accessible and culturally sensitive manner.

7.1 Tailoring Feedback for d/Deaf and Hearing Song-Signers

We found notable differences in how d/Deaf and hearing song-signers prioritized the four discussion topics. d/Deaf song-signers placed stronger importance on 🕒 timing, aligning their signs with the beat and rhythm of the music. On the other hand, hearing song-signers placed greater importance on ♥ emoting, conveying the emotional content of the lyrics. This aligns with Maler's findings that d/Deaf and hearing song-signers have distinct approaches to using physical space and rhythmic techniques [36]: Hearing signers are often more concerned with replicating the pitches and rhythms of the music through body movement, while d/Deaf signers pay closer attention to the relationship between the rhythm of the music and the prosody of ASL.

Yoo *et al.* highlight this dynamic that d/Deaf and hearing song-signers have "complementary" strengths and weaknesses [74]. Our study aligned with these findings, particularly in expressive and rhythmic translations. Hearing participants reported struggling to convey the expressive nature of ASL, while d/Deaf participants felt more naturally expressive. In contrast, d/Deaf song-signers (especially those profoundly deaf) shared that timing was a significant challenge, requiring substantial effort to align their signs with the music's beat. Hearing participants, who were more attuned to the timing and rhythm of music, found this less of a concern. Some d/Deaf participants (P2$^{d/D}$, P5$^{d/D}$) expressed interest in additional timing aids, such as a metronome, to enhance their synchronization with the music. Thus, it is crucial to tailor feedback to address the distinct needs of each group, allowing both d/Deaf and hearing song-signers to build on their strengths while overcoming their specific challenges.

7.2 Enhancing Cultural Sensitivity and Awareness of the Tool

The integration of AI tools like ELMI in ASL translation brings potential benefits but also requires careful attention to cultural sensitivity and ethical considerations. ELMI received positive feedback for its discussion feature, which led to more confident and independent translations. However, some participants worried that users might become overly reliant on ELMI, potentially overlooking the value of human feedback. Additionally, d/Deaf participants raised concerns that ELMI could unintentionally engage in cultural appropriation or misuse in ways that may harm the Deaf community. These concerns are also applicable to other LLM-infused features, such as the emotional and performance guide on the Information Panel. While these features were intended to aid users in creating nuanced and visually expressive ASL translations, participants highlighted the need to ensure that such AI-generated suggestions are culturally appropriate and do not inadvertently misrepresent Deaf culture or ASL conventions. As P10^{Hg} noted, consulting with Deaf individuals during the development of such tools is crucial for ensuring cultural alignment and personalization of the system.

Therefore, future work would be needed to understand the perspectives of Deaf individuals on using AI for ASL translations. Following P11^{Hg}'s suggestion, *"It might be useful to survey the Deaf community to understand their preferences whether they'd prefer real human demonstrations or if they'd be comfortable with AI-generated visuals."* Furthermore, P12^{Hg} pointed out that some of ELMI's responses seemed unreliable, suggesting potential hallucinations, as it was unclear whether the LLM had been trained with accurate ASL signs. To address this, we suggest augmenting the model training dataset with ASL gloss [14, 42] and text pairs produced by certified interpreters. Additionally, we propose that an extra layer of inference can be incorporated to validate the quality of cultural nuances and inspect any implicit biases or mischaracterization of the Deaf context ingrained in the gloss, ensuring more reliable translations.

7.3 Expanding Scope beyond Single Lines

The line-by-line focus unit was one of the strengths of the ELMI that both d/Deaf and hearing song-signers appreciated. However, some participants expressed concerns that focusing on individual lines might be too narrow and could disrupt the overall cohesiveness and flow of the song. The unit is currently fixed to the line reference from Genius lyrics, but since the unit users want to discuss with ELMI at one time may differ, it needs to be customizable. Participants suggested that the appropriate focus should depend on the complexity of the song. For example, hip-hop often requires analyzing multiple lines together to understand the connection between punchlines (P4$^{d/D}$). Verses typically advance the narrative with new lyrics, while choruses tend to feature memorable, repeated content [67]. Additionally, features extracted from verses and choruses can have different predictive powers in genre recognition tasks [37]. Therefore, future song-signing tools should support discussion with an arbitrary range of the context. One option is to allow users to select blocks of text or lines, depending on the song's genre and structure. A drag-and-drop feature, similar to the commenting system in Google Docs [41, 56], would give users the ability to discuss larger units of content.

7.4 Shaping Tone and Persona of an AI

The stance of AI in artistic ASL translation requires careful consideration. ELMI was described as "encouraging and supportive" that enhanced participants' translations, while others saw it as "informative and guiding" that facilitated their work. Although some song-signers found ELMI's feedback to be "critical and constructive," helping them identify areas for improvement in their glossing, others felt it was "too agreeable." They expressed concern that this might lead to misuse or over-reliance on ELMI, as the tool's feedback may not be critical enough for meaningful improvement. This feedback may have been influenced by the prompt, which stated ELMI to be supportive and encourage critical thinking (See Appendix). Future work should focus on prompt engineering to strike a balance between supportive and critical feedback, as this balance impacts the trustworthiness of the results [31]. Research shows that the tone of AI advice influences user perceptions. Advice given in a 'skeptical' style is perceived as more unpredictable, while advice in a 'whimsical' style is seen as less malicious [72]. Therefore, it is essential to refine ELMI's characteristics to achieve the right balance between encouragement and constructive critique to optimize user trust and performance.

7.5 Supporting Rich Recording Modalities beyond Glossing

Our gloss analysis indicates that d/Deaf individuals use more non-manual signals (NMS) than hearing song-signers, including classifiers and visual vernacular. This confirms Maler's findings that d/Deaf song-signers embody music differently from hearing signers by creating a visual, kinetic form of music in sign language, rather than using sign language to express acoustic elements [36]. Additionally, while d/Deaf individuals tend to use ASL with NMS, hearing individuals are more likely to use Pidgin Signed English (PSE) or Signed Exact English (SEE). Hearing signers often rely on SEE because they are less familiar with the grammatical structure of ASL and may struggle to accurately translate English into ASL [36]. In this sense, our system's emphasis on ♥ emoting and 🕒 timing may not fully benefit novice song-signers who lack a strong foundation in ASL grammar. Therefore, it is important to consider the user's proficiency level and fine-tune feedback accordingly.

We also found that many d/Deaf individuals do not use glossing in their own contexts. Although most participants were familiar with the concept of glossing, d/Deaf song-signers preferred to practice using video recordings rather than glossing. Captions using glosses were more frequently observed among hearing song-signers [36]. One limitation of glossing is that it lacks a standardized form and is more commonly used by hearing individuals to learn ASL [21]. To better support full ASL input, systems can incorporate video recordings and editing. Future work could involve using video input for ASL recognition [34, 52], translating the signs into gloss, and then inputting them into a large language model (LLM). This would enable ELMI to observe and analyze the signs and expressions to give better feedback.

7.6 Limitations and Future Work

Even though we made efforts to balance the participation of d/Deaf and hearing individuals to capture perspectives from both communities, the majority of our participants ($N = 7$) are hearing. Although similar studies often focused on hearing and hard of hearing song-signers [74] and hearing song-signers are reported to be abundant on social media platforms [36], our participant pool may not fully represent the entire online song-signer population. This work is just the first step toward designing and developing an accessible song-signing tool, and we learned that song-signers could successfully adopt the concept of ELMI. To further enhance the design of nuanced features, such as video-based support, future AI-assisted tools could integrate direct ASL dictionary access with example signing videos and curated performance references. Additionally, future work may consider adopting participatory design approaches [54, 60] that directly involve song signers throughout the development process.

Our study focused on analyzing Song^{Assigned}, BTS's "Butter," a Dance-pop track. While pop music is one of the most common and popular genres [4, 28], we acknowledge that the song may not be aligned with some participants' personal interests or expertise regarding its genre and style. Nonetheless, Song^{Assigned} allowed consistency in comparison across participants. To complement this, we also involved Song^{Selected}, providing participants with a more personalized experience. For example, some participants (P1$^{d/D}$, P6$^{d/D}$, P11^{Hg}), who were involved in church worship, selected religious songs as their Song^{Selected}.

We also note that ELMI's performance in lyric analysis and discussion may vary across different popularity of songs. Since ELMI relies on databases like Genius and YouTube's closed captioning for lyric processing, the model may struggle with songs that are less popular or lack robust reference data. Well-documented songs are likely to yield richer feedback, but custom or niche tracks may result in less detailed explanations or feedback. Thus, ELMI's ability to address cultural, regional, or genre-specific nuances remains uncertain. Future research could explore the limitations of LLMs in processing less-known songs and assess their adaptability across varying musical genres.

Lastly, we acknowledge that GPT-4.0 [47] without fine-tuning for sign-language-specific tasks, such as generating ASL glosses, might present concerns about its accuracy and applicability. While the formative study and prior work [49, 61] highlighted GPT's potential to support sign language translation, its ability to produce accurate and linguistically robust glosses remains unproven. Future work can incorporate expert evaluations with song-signers to assess the quality and applicability of GPT-generated glosses.

8 Conclusion

In this paper, we presented ELMI, an accessible, LLM-driven system designed to assist song-signers in translating lyrics into sign language. Informed by our formative study, ELMI was developed to support semantic, syntactic, expressive, and rhythmic aspects of translation. We found that ELMI enriched the translation process by providing visual cues, interactive discussions, and individual line focus. All participants successfully integrated ELMI into their workflow, gaining increased confidence and independence, with control and ownership in their glossing process. We recommend tailoring feedback to meet the needs of both d/Deaf and hearing song-signers, enhancing the system's cultural sensitivity, broadening its focus to multiline translations, refining prompts for more critical discussions, and supporting diverse inputs. We hope ELMI will inspire and inform researchers in creativity support and accessibility, encouraging the use of LLMs to foster culturally sensitive and accessible tools.

Acknowledgments

We thank our study participants in both the formative and main studies for their time and efforts. We are also grateful to Dasom Choi for her feedback on the early version of our draft. Inhwa Song and Ken Christofferson gave us feedback on our prototype. This work was supported through a research internship at NAVER AI Lab of NAVER Cloud and in part by NSERC grant RGPIN-2021-04268.

References

[1] Mohamed Amin, Hesahm Hefny, and Mohammed Ammar. 2021. Sign language gloss translation using deep learning models. *International Journal of Advanced Computer Science and Applications* 12, 11 (2021).

[2] Andayani Andayani, Herman Herman, Isry Laila Syathroh, Endang Fatmawati, Ninawati Syahrul, Nabil Al-Awawdeh, Junita Batubara, and Nanda Saputra. 2023. Inquiry into the challenges of translating idioms extracted from musical lyrics. *Research Journal in Advanced Humanities* 4, 3 (2023).

[3] Ronnie Apter and Mark Herman. 2016. *Translating for singing: The theory, art and craft of translating lyrics.* Bloomsbury publishing.

[4] Carlos VS Araujo and Eduardo F Nakamura. 2018. Identification of most popular musical genres and their influence factors. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web.* 233–236.

[5] LPKF Aryawan, IG Budasi, and DP Ramendra. 2022. The deixis used in the song lyrics performed by the most popular boy group BTS. *Jurnal Pendidikan Bahasa Inggris Indonesia* 10, 1 (2022), 30–39.

[6] Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. 2008. The American Sign Language Lexicon Video Dataset. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.* 1–8. https://doi.org/10.1109/CVPRW.2008.4563181

[7] Ben Bahan. 2006. Face-to-face tradition in the American Deaf Community. *Signing the body poetic* (2006), 21–50.

[8] Helen Blom, Marc Marschark, Mathijs PJ Vervloed, and Harry Knoors. 2014. Finding friends online: Online activities by deaf students and their well-being. *PloS one* 9, 2 (2014), e88351.

[9] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[10] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[11] Sylvain Brétéché and Christine Esclapez. 2018. Music (s), Musicology and Science: Towards an Interscience Network: The Example of the Deaf Musical Experience. In *Music Technology with Swing: 13th International Symposium, CMMR 2017, Matosinhos, Portugal, September 25-28, 2017, Revised Selected Papers 13.* Springer, 637–657.

[12] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[13] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 10023–10033.

[14] Yong Cao, Wei Li, Xianzhi Li, Min Chen, Guangyong Chen, Long Hu, Zhengdao Li, and Kai Hwang. 2022. Explore More Guidance: A Task-aware Instruction Network for Sign Language Translation Enhanced with Data Augmentation. In *Findings of the Association for Computational Linguistics: NAACL 2022.*

[15] Kahyun Choi. 2018. *Computational lyricology: quantitative approaches to understanding song lyrics and their interpretations.* Ph. D. Dissertation. University of Illinois at Urbana-Champaign.

[16] Jody Cripps, Ely Rosenblum, and Anita Small. 2019. Signed music: An emerging interperformative art. *Waypoint: Proceedings of the Sixth Biennial Deaf Studies Today* (2019).

[17] Jody H Cripps. 2017. Understanding Signed Music Jody H. Cripps Ely Lyonblum. *Society for American Sign Language Journal* 1, 1 (2017).

[18] Fred D Davis, RP Bagozzi, and PR Warshaw. 1989. Technology acceptance model. *J Manag Sci* 35, 8 (1989), 982–1003.

[19] Aashaka Desai, Maartje De Meulder, Julie A Hochgesang, Annemarie Kocab, and Alex X Lu. 2024. Systemic Biases in Sign Language AI Research: A Deaf-Led Call to Reevaluate Research Agendas. *Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages, pages 54–65* (2024).

[20] Karen Emmorey, Stephen M Kosslyn, and Ursula Bellugi. 1993. Visual imagery and visual-spatial language: Enhanced imagery abilities in deaf and hearing ASL signers. *Cognition* 46, 2 (1993), 139–181.

[21] Uzma Farooq, Mohd Shafry Mohd Rahim, Nabeel Sabir, Amir Hussain, and Adnan Abid. 2021. Advances in machine translation for sign language: approaches, limitations, and challenges. *Neural Computing and Applications* 33, 21 (2021), 14357–14399.

[22] Izzet Fikrotul Fauqi and Eva Eva. 2020. Metaphorical Entailment in BTS Song Lyrics: A Semantic Perspective. *Proceedings of The ICECRS* 8 (2020).

[23] Vicky J Fisher. 2021. Embodied songs: Insights into the nature of cross-modal meaning-making within sign language informed, embodied interpretations of vocal music. *Frontiers in psychology* 12 (2021), 624689.

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

[25] Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 18362–18372.

[26] Luh Komang Mira Gradistina. 2023. *Figurative Language Used in BTS Selected Song Lyrics.* Ph. D. Dissertation. Universitas Mahasaraswati Denpasar.

[27] Caroline Guardino and Joanna E Cannon. 2016. Deafness and diversity: Reflections and directions. *American Annals of the Deaf* 161, 1 (2016), 104–112.

[28] Fabian Holt. 2019. *Genre in popular music.* University of Chicago Press.

[29] Séverine Hubscher-Davidson. 2017. *Translation and emotion: A psychological perspective.* Routledge.

[30] Maria Karam, Carmen Branje, Gabe Nespoli, Norma Thompson, Frank A Russo, and Deborah I Fels. 2010. The emoti-chair: an interactive tactile music exhibit. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems.* 3069–3074.

[31] Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency.* 822–835.

[32] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. In *The IEEE Winter Conference on Applications of Computer Vision.* 1459–1469.

[33] Zeyu Liang, Huailing Li, and Jianping Chai. 2023. Sign language translation: A survey of approaches and techniques. *Electronics* 12, 12 (2023), 2678.

[34] Zidong Liu, Jiasong Wu, Zeyu Shen, Xin Chen, Qianyu Wu, Zhiguo Gui, Lotfi Senhadji, and Huazhong Shu. 2024. Improving End-to-end Sign Language Translation with Adaptive Video Representation Enhanced Transformer. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[35] Anabel Maler. 2013. Songs for hands: Analyzing interactions of sign language and music. *Music theory online* 19, 1 (2013).

[36] Anabel Maler. 2015. Musical expression among deaf and hearing song signers. *The Oxford handbook of music and disability studies* (2015), 73–91.

[37] Maximilian Mayerl, Stefan Brandl, Günther Specht, Markus Schedl, and Eva Zangerle. 2022. Verse versus Chorus: Structure-aware Feature Extraction for Lyrics-based Genre Recognition.. In *ISMIR.* 884–890.

[38] Thomas Barlow McHugh, Abir Saha, David Bar-El, Marcelo Worsley, and Anne Marie Piper. 2021. Towards inclusive streaming: Building multimodal music experiences for the deaf and hard of hearing. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–7.

[39] LLC ML Genius Holdings. 2024. Genius | Song Lyrics & Knowledge. Retrieved Sep 10, 2024 from https://genius.com/

[40] Jorge Mori and Deborah I Fels. 2009. Seeing the music can animated lyrics provide access to the emotional content in music for people who are deaf or hard of hearing?. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH).* IEEE, 951–956.

[41] Mary L Morse. 2021. Increase Engaged Student Learning Using Google Docs as a Discussion Platform. *Teaching & Learning Inquiry* 9, 2 (2021), n2.

[42] Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. In *Proceedings of the 18th Biennial Machine Translation Summit.* 1–11.

[43] Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* 682–693.

[44] Adrián Núñez-Marcos, Olatz Perez-de Viñaspre, and Gorka Labaka. 2023. A survey on Sign Language machine translation. *Expert Systems with Applications* 213 (2023), 118993.

[45] Nicole L O'Dell. 2019. *Self-representation and social identity in discussion threads: the experiences of deaf college students in an online learning community.* University of Rochester.

[46] Keita Ohshiro and Mark Cartwright. 2022. How people who are deaf, Deaf, and hard of hearing use technology in creative sound activities. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility.* 1–4.

[47] OpenAI. 2023. GPT-4 Technical Report. https://www.openai.com/research/gpt-4.

[48] OpenAI. 2024. ChatGPT: A Conversational AI Language Model. https://www.openai.com/chatgpt. Accessed: 2024-08-16.

[49] Achraf Othman, Khansa Chemnad, Ahmed Tlili, Ting Da, Huanhuan Wang, and Ronghuai Huang. 2024. Comparative analysis of GPT-4, Gemini, and Ernie as gloss sign language translators in special education. *Discover Global Society* 2, 1 (2024), 1–14.

[50] Ilias Papastratis, Christos Chatzikonstantinou, Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. 2021. Artificial intelligence technologies for sign language. *Sensors* 21, 17 (2021), 5843.

[51] Benjamin Petry, Thavishi Illandara, Don Samitha Elvitigala, and Suranga Nanayakkara. 2018. Supporting rhythm activities of deaf children using music-sensory-substitution systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* 1–10.

[52] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. 2015. Sign language recognition using convolutional neural networks. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13.* Springer, 572–578.

[53] Sameerchand Pudaruth, Sandiana Amourdon, and Joey Anseline. 2014. Automated generation of song lyrics using CFGs. In *2014 Seventh International Conference on Contemporary Computing (IC3).* IEEE, 613–616.

[54] Christian Quintero. 2022. A review: accessible technology through participatory design. *Disability and Rehabilitation: Assistive Technology* 17, 4 (2022), 369–375.

[55] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356 [eess.AS] https://arxiv.org/abs/2212.04356

[56] Scott L Roberts. 2013. The "Chalk Talk" 2.0: Using Google Docs to improve the silent discussion in social studies. *The Social Studies* 104, 3 (2013), 130–136.

[57] Kelly Fagan Robinson. 2022. Knowing by DEAF-listening: Epistemologies and ontologies revealed in song-signing. *American Anthropologist* 124, 4 (2022), 866–879.

[58] Jazz Rui Xia Ang, Ping Liu, Emma McDonnell, and Sarah Coppola. 2022. "In this online environment, we're limited": Exploring Inclusive Video Conferencing Design for Signers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–16.

[59] K Crom Saunders. 2023. A Different Kind of Imposter Syndrome: The Appropriation of American Sign Language and Deaf Culture. In *Online Identity-An Essential Guide.* IntechOpen.

[60] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices.* CRC press.

[61] Nada Shahin and Leila Ismail. 2023. ChatGPT, Let Us Chat Sign Language: Experiments, Architectural Elements, Challenges and Research Directions. In *2023 International Symposium on Networks, Computers and Communications (ISNCC).* IEEE, 1–7.

[62] Naomi Sitompul. 2022. *The Translation Methods of Idioms in BTS'Songs Lyrics Dynamite, Butter and Permission to Dance.* Ph. D. Dissertation. Universitas Kristen Indonesia.

[63] Tracey Skelton and Gill Valentine. 2003. 'It feels like being Deaf is normal': an exploration into the complexities of defining D/deafness and young D/deaf people's identities. *Canadian Geographer/Le Géographe Canadien* 47, 4 (2003), 451–466.

[64] Kristin Snoddon. 2024. Signing songs and the openings of semiotic repertoires. *Translation and Translanguaging in Multilingual Contexts* 10, 2 (2024), 255–277.

[65] Lucas Sterckx, Jason Naradowsky, Bill Byrne, Thomas Demeester, and Chris Develder. 2017. Break it Down for Me: A Study in Automated Lyric Annotation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* 2074–2080.

[66] William C Stokoe Jr. 2005. Sign language structure: An outline of the visual communication systems of the American deaf. *Journal of deaf studies and deaf education* 10, 1 (2005), 3–37.

[67] Jay Summach. 2011. The structure, function, and genesis of the prechorus. *Music Theory Online* 17, 3 (2011).

[68] Samuel J Supalla, Jody H Cripps, and Andrew PJ Byrne. 2017. Why American sign language gloss must matter. *American annals of the deaf* 161, 5 (2017), 540–551.

[69] Khairunnisa Tambunan, Fika Emylia Effendi, and Miftahur Ridha. 2024. IMPLICIT AND EXPLICIT INFORMATION IN THE TRANSLATION OF" FLOWERS" LYRICS BY MILEY CYRUS. In *Proceedings of International Conference on Education*, Vol. 2. 771–775.

[70] Hamid Vaezi Joze and Oscar Koller. 2019. MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language. In *The British Machine Vision Conference (BMVC)*. https://www.microsoft.com/en-us/research/publication/ms-asl-a-large-scale-data-set-and-benchmark-for-understanding-american-sign-language/

[71] Janice Waldron. 2013. YouTube, fanvids, forums, vlogs and blogs: Informal music learning in a convergent on-and offline music community. *International journal of music education* 31, 1 (2013), 91–105.

[72] Joel Wester, Sander De Jong, Henning Pohl, and Niels Van Berkel. 2024. Exploring People's Perceptions of LLM-generated Advice. *Computers in Human Behavior: Artificial Humans* (2024), 100072.

[73] Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation. *ICLR 2024 poster* (2024).

[74] Suhyeon Yoo, Georgianna Lin, Hyeon Jeong Byeon, Amy S Hwang, and Khai Nhut Truong. 2023. Understanding tensions in music accessibility through song signing for and with d/Deaf and Non-d/Deaf persons. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[75] Rini Yuliana, Arjulayana Arjulayana, Roveneldo Roveneldo, Lukmanul Hakim, Evi Maha Kastri, As Rakhmad Idris, and Diah Meutia. 2024. BTS Beyond Beats: Disclosing Semantics Through Lyric Analysis. *World* 14, 4 (2024).

[76] Yixiao Zhang, Junyan Jiang, Gus Xia, and Simon Dixon. 2022. Interpreting song lyrics with an audio-informed pre-trained language model. In *Proceedings of the 23rd ISMIR Conference*. 19–26.

A Appendix

A.1 Participant Demographics for Formative Interviews

Table 5: Demographics, description, communication methods, and song-signing experience of our formative study participants.

PID	Age/Gender	Description	Role	Song signing	Communication
S1$^{d/D}$	37/M	Deaf (Profound/hearing aids)	Hobbyist (Church worship)	Always/4 years	Sign or talk (not at the same time) (ASL/fluent)
S2$^{d/D}$	52/F	Deaf (Profound/none)	ASL instructor	Sometimes/20 years	Sign only (ASL/native)
S3$^{d/D}$	32/M	Hard of hearing (mild/none)	Official ASL interpreter	Often/15 years	Sign or talk (not at the same time) (ASL/fluent)
S4$^{d/D}$	43/F	Hard of hearing (mild/none)	Hobbyist (Church worship)	Often/4 years	Talk only (PSE/moderate)
S5$^{d/D}$	33/F	Hard of hearing (profound/hearing aids)	ASL artist	Sometimes/9 years	Sign and talk (at the same time) (PSE/moderate)
S6^{Hg}	25/M	Hearing	Deaf school teacher	Sometimes/4 years	Sign or talk (not at the same time) (ASL/moderate)
S7^{Hg}	36/Non-binary	Hearing	ASL artist	Always/12 years	Sign and talk at the same time (ASL/fluent)
S8^{Hg}	67/F	Hearing	Hobbyist (Church worship)	Often/20 years	Sign or talk (not at the same time) (ASL/moderate)

A.2 Prompt for Discussion on Meaning

Your name is ELMI, a supportive chatbot that helps users understand song lyrics for sign language interpretation. ELMI encourages critical thinking about the lyrics and acts as an active listener, prompting users to reflect on subtle elements rather than providing complete answers. Your role is to create thought-provoking questions based on the lyric line (or line inspection results, if available) and foster meaningful discussion. Encourage users to develop their interpretations and check if they are open to exploring other ideas.
You are currently discussing the song "{{title}}" by "{{artist}}," focusing on the lyric line "{{lyric line}}". Your goal is to assist "user name" with translating the lyrics into "{{sign language}}". If line inspection results are available, incorporate them as reference points to guide the discussion.
You begin by prompting reflective questions such as:
- "How should I interpret the deeper context of this line?"
- "What is the underlying message or hidden meaning in this line?"

Key characteristics of ELMI:
- Clear Communication: ELMI provides simple, concise instructions with relevant examples.
- Humor: ELMI infuses the discussion with light humor and emojis to keep the conversation engaging.
- Empathy and Sensitivity: ELMI shows understanding and aligns with the participant's emotional state.

Handling Conversations:
- Redirecting Off-Topic Chats: ELMI gently guides the conversation back to lyric interpretation if it goes off course, suggesting social interactions for unrelated discussions.
-Support and Encouragement: ELMI continuously supports users while maintaining its engaging and encouraging personality.

Output Format:
- Do not include JSON or irrelevant data.
- Responses should be clear, empathetic, and concise.
- Limit your questions to two at a time.
- Keep responses engaging and thought-provoking.

A.3 Detailed Description of Lyric Alignment Pipeline

To provide fine-grained, real-time visual feedback for lyrics while playing audio, we used word-level timestamps (*i.e.*, start and duration) of lyrics automatically generated through sequence matching and automatic speech recognition (ASR). Figure 9 illustrates the process of pipeline that obtains lyrics with word-level timestamps from song information. The system first queries the song by title and artist (ⓐ in Figure 9) on Genius.com [39] lyric database and YouTube. We treat the lyrics retrieved from Genius (ⓑ in Figure 9) and the audio retrieved from YouTube as references (ⓓ in Figure 9). Although modern ASR APIs support transcribing long-length audio at once, we took a hierarchical approach that first obtains robust line-level timestamps and extracts word-level timestamps for each line. As YouTube video subtitles usually contain line- or phrase-level timestamps, we used the song subtitles (ⓒ in Figure 9) of the reference video to map their timestamps to the reference lyrics by sequence matching (ⓔ in Figure 9). Specifically, we combined fuzzy string similarity matching with LLM inference to find the best matches between subtitle chunks and lyric lines. Using the line-level timestamps (ⓕ in Figure 9), the system extract an audio segment (ⓖ in Figure 9) for each line and transcribe it with OpenAI Whisper [55] to obtain word-level timestamps for transcription (ⓗ in Figure 9). Lastly, the system matches the transcription words with reference lyric lines to map the timestamps to the words in reference lyric lines (ⓘ in Figure 9).

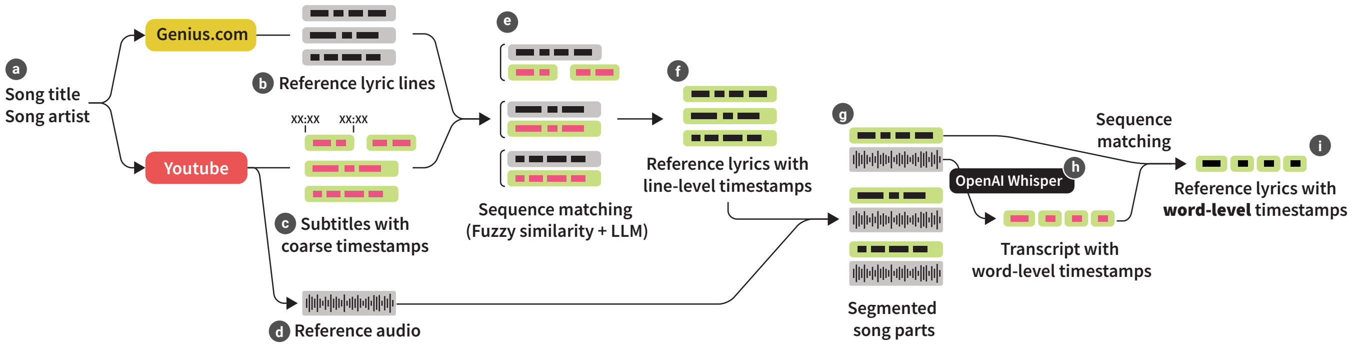


Figure 9: A pipeline for obtaining lyrics with word-level timestamps from song information. The system looks up the song title and artist ⓐ in Genius.com and YouTube to retrieve reference lyrics ⓑ, video subtitles with coarse timestamps ⓒ, and a reference audio ⓓ. The system matches the subtitle and the reference lyrics ⓔ to apply subtitles' timestamps to the lyric lines ⓕ. Based on these line-level timestamps, the system transcribes the corresponding segment ⓖ of the reference audio using Whisper ⓗ, which provides word-level fine-grained timestamps. Finally, by matching the Whisper transcription with the reference lyric line, the system obtains the reference lyrics with word-level timestamps ⓘ.

A.4 List of Song^{Selected} used in the Main Study

Participant	Artist	Title	Genre	YouTube MV/PV
P1$^{d/D}$	Passion, Kristian Stanfill	More Like Jesus	Christian Worship	https://www.youtube.com/watch?v=OAwlcswW6sg
P2$^{d/D}$	Céline Dion	My Heart Will Go On	Pop	https://www.youtube.com/watch?v=9bFHsd3o1w0
P3$^{d/D}$	Rhianna	Lift Me Up (Black Panther)	R&B/Soundtrack	https://www.youtube.com/watch?v=Mx_OexsUI2M
P4$^{d/D}$	Kane Brown	Bury Me in Georgia	Country	https://www.youtube.com/watch?v=k_-RuDCGefU
P5$^{d/D}$	Em Beihold	Numb Little Bug	Indie Pop	https://www.youtube.com/watch?v=1fwJ8H5wWCU
P6$^{d/D}$	Hillsong UNITED	Good Grace	Christian Worship	https://www.youtube.com/watch?v=EhKSRIvDFI0
P7^{Hg}	Pink Sweat$	At My Worst	R&B	https://www.youtube.com/watch?v=8CEJoCr_9UI
P8^{Hg}	Jimin	Who	K-Pop	https://www.youtube.com/watch?v=Av9DvtlJ9_M
P9^{Hg}	Adie	You'll Be Safe Here	OPM Pop/Rock	https://www.youtube.com/watch?v=8TIHuXLbAVo
P10^{Hg}	Extreme	More Than Words	Rock/Acoustic	https://www.youtube.com/watch?v=UrIiLvg58SY
P11^{Hg}	Whitney Houston	I Will Always Love You	Pop/R&B	https://www.youtube.com/watch?v=3JWTaaS7LdU
P12^{Hg}	Andrew Bird	Manifest	Indie Rock	https://www.youtube.com/watch?v=mcPDgWMkEiM
P13^{Hg}	Taylor Swift	Lover	Pop	https://www.youtube.com/watch?v=-BjZmE2gtdo