

Towards AI-driven Sign Language Generation with Non-manual Markers

Han Zhang*
University of Washington, USA
micohan@cs.washington.edu

Rotem Shalev-Arkushin*
Tel-Aviv University, Israel
rotems7@mail.tau.ac.il

Vasileios Baltatzis
Apple, USA
vbaltatzis@apple.com

Connor Gillis
Apple, USA
connorgillis@apple.com

Gierad Laput
Apple, USA
gierad@apple.com

Raja Kushalnagar*
Gallaudet University, USA
raja.kushalnagar@gallaudet.edu

Lorna Quandt*
Gallaudet University, USA
lorna.quandt@gallaudet.edu

Leah Findlater
Apple, USA
lfindlater@apple.com

Abdelkareem Bedri
Apple, USA
bedri@apple.com

Colin Lea
Apple, USA
colin_lea@apple.com

"Do you commute to work by bike?"

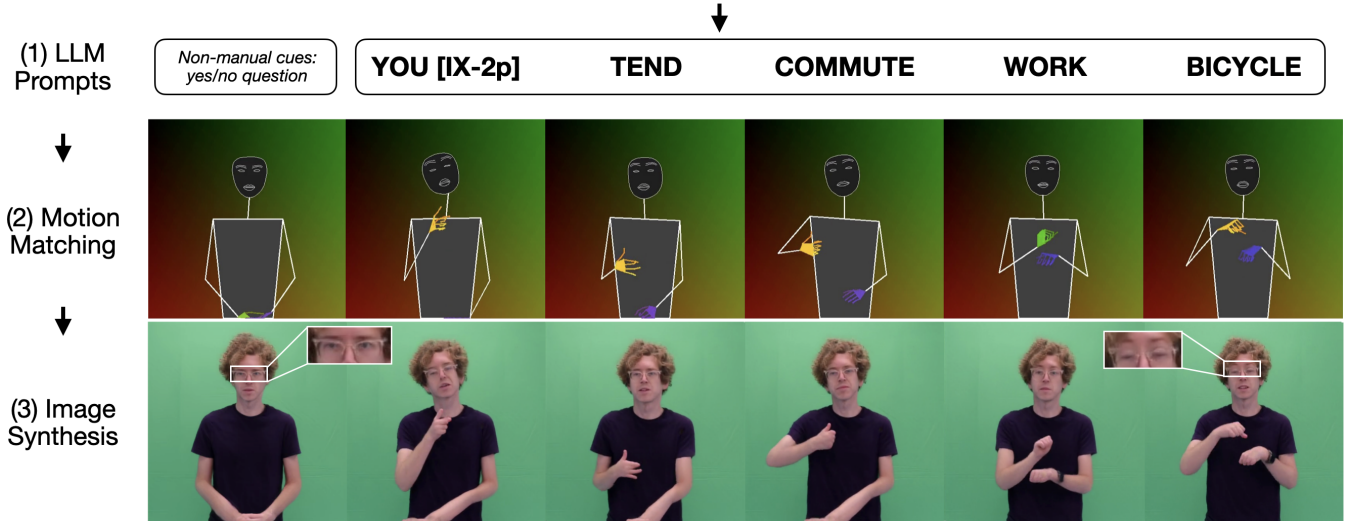


Figure 1: Our prototype translates English text into a photorealistic ASL video which includes both manual and non-manual information. It starts with an English text input (top), and translates it into ASL representations capturing both manual elements (e.g., hand movements) and non-manual information (e.g., facial expressions). From those, it produces a skeletal pose sequence, and finally converts it into a photorealistic ASL video. In this example, raised eyebrows signal a yes/no question. Without this non-manual marker, the same sentence would be interpreted as a statement.

*Work done entirely at Apple.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

CHI '25, Yokohama, Japan

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

ACM ISBN 979-8-4007-1394-1/25/04

<https://doi.org/10.1145/3706598.3713855>

Abstract

Sign languages are essential for the Deaf and Hard-of-Hearing (DHH) community. Sign language generation systems have the potential to support communication by translating from written languages, such as English, into signed videos. However, current systems often fail to meet user needs due to poor translation of grammatical structures, the absence of facial cues and body language, and insufficient visual and motion fidelity. We address these

challenges by building on recent advances in LLMs and video generation models to translate English sentences into natural-looking AI ASL signers. The text component of our model extracts information for manual and non-manual components of ASL, which are used to synthesize skeletal pose sequences and corresponding video frames. Our findings from a user study with 30 DHH participants and thorough technical evaluations demonstrate significant progress and identify critical areas necessary to meet user needs.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods; Accessibility systems and tools.**

Keywords

Sign language generation, assistive technology, accessibility, human-centered design, DHH community

ACM Reference Format:

Han Zhang, Rotem Shalev-Arkushin, Vasileios Baltatzis, Connor Gillis, Gierad Laput, Raja Kushalnagar, Lorna Quandt, Leah Findlater, Abdelkareem Bedri, and Colin Lea. 2025. Towards AI-driven Sign Language Generation with Non-manual Markers. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26–May 1, 2025, Yokohama, Japan. ACM, New York, NY, USA, 26 pages. <https://doi.org/10.1145/3706598.3713855>

1 Introduction

Sign languages are crucial for communication within the Deaf and Hard-of-Hearing (DHH) communities [46, 112]. As naturally-emerging and fully-fledged languages, they enable individuals to convey complex ideas, emotions, and cultural nuances through movements and facial expressions [37, 124]. Despite their importance for many DHH people, communication barriers between signing and non-signing communities exist due to limited access to skilled sign language interpreters, low levels of sign language proficiency among the general population, and the exclusion of sign languages from most communication technologies designed for spoken or written languages [16, 92, 99].

Systems that translate spoken languages into sign languages (sign language generation, SLG) and vice versa (sign language translation, SLT) hold promise to bridge this communication gap [124, 138]. In this work we focus on the SLG task, specifically translating from English text to American Sign Language (ASL). Historically, SLG technology has faced criticism from DHH users due to low-fidelity avatars, poor language translation, and oversimplification of sign linguistics [65, 66, 77]. Recent research, however, has suggested that improved quality and overcoming technical limitations could increase acceptance among DHH individuals [71, 120]. Our work uses advances in machine learning (ML) to develop an SLG prototype system and investigate whether technological improvements meet the needs and interests of the DHH and signing community.

Sign languages combine manual markers—such as hand movements, orientation, and location—with non-manual markers, including facial expressions, head movements, and other body language, to create grammatical structures and convey meaning [17, 127, 136]. For example, in ASL, manual markers such as location and movement within the signing space can modify a sign’s grammatical

function, indicating subjects, objects, or other syntactic roles [136]. Similarly, non-manual markers can also indicate critical information, such as a head shake accompanying a sign to denote negation, raised eyebrows and a distinct facial expression to form conditional clauses or emphasize the topic of a sentence or raised eyebrows and a forward head tilt to signal a yes/no question [7, 9, 127], as exemplified in Figure 1. Each of these linguistic aspects of ASL presents a challenge for modern SLG systems, given that natural, understandable signing must include sufficient information shown in a fluid manner to convey multiple distinct streams of information.

While recent work on SLG has progressed [40, 57, 64, 94, 130, 137], these systems typically take a generic view of signing, often overlooking sign language nuances, including the role of non-manual markers. To address these challenges, we prototype a modular ASL generation system designed to produce automated signing by simultaneously focusing on technical improvements, user perceptions, and the unique linguistic structure of ASL. Our system is tailored for open-ended, context-free use cases, allowing users to input an English sentence and generate a signed video that appears natural and comprehensive. Developing an effective SLG system capable of modeling complex signed interactions is a grand challenge that requires interdisciplinary expertise, alongside stewardship from the DHH and signing community [16]. Guided by this principle, our research prototype was developed and refined through collaboration among researchers from diverse fields, including those in computer vision, computer graphics, human-computer interaction, and experts from the DHH and signing communities. It consists of three modules: (1) translating English text into intermediate ASL representations—including English-based glosses to capture manual markers and linguistic information to represent non-manual markers—using few-shot approach with GPT-4o, (2) synthesizing human pose and body motions from these representations using a Motion Matching approach, and (3) generating photorealistic signed video frames representing an ASL signer using an image generation model.

We conducted both technical evaluations and a user study with 30 DHH signers to assess our prototype system and to gauge the interest of DHH individuals in its use. The technical evaluation examined the translation of English sentences into ASL written representations, including manual and non-manual components, and the generation of signed videos. The user study evaluated translation quality, visual fidelity, and motion naturalness, while gathering perspectives on potential use cases. Our findings indicate that the system achieves compelling translation performance relative to reported results in the literature. However, there remains significant room for improvement. While participants were frequently able to understand the content of the signed videos, their perceptions on the signing quality, particularly in comparison to real human signers, were less favorable.

In summary, our contributions include:

- In Section 3, we introduce a modular ASL generation prototype designed to produce natural and comprehensive signed videos that includes non-manual cues.
- In Section 4, we present technical evaluations of our approach. Results show a BLEU-4 score of 0.276 for English Text-to-ASL gloss translation, an average precision of 0.91

and recall of 0.97 for detecting non-manual information from English text, and improved video generation performance over baseline methods.

- In Section 5, we detail a user study assessing the perceived translation quality of our system, as well as the visual and motion quality of its outputs. Results indicate both potential and tangible areas for improvement, alongside insights into the system’s potential use cases (*e.g.*, doctor office and video or in-person conversations).
- In Section 6, we reflect on our design process, share key insights gained from our design and evaluation process, provide recommendations on how to address remaining challenges, and discuss computational and ethical considerations in the use of our system.

While continued effort is needed to advance SLG systems in collaboration with the DHH and signing communities, our work represents an initial step in addressing critical technical challenges and taking a comprehensive approach to ASL. It demonstrates the potential of these systems and encourages further exploration of critical aspects of signing, especially non-manual markers.

2 Background and Related Work

In this section, we overview Deaf cultures and sign languages, review sign language generation systems, focusing on their technical challenges, and discuss the DHH users’ perspectives on sign language technologies.

2.1 Deaf Cultures and Sign Languages

In 1970, the term “Deaf Culture” was developed to articulate that many Deaf communities possess their own ways of life, characterized by a shared set of values, behaviors, traditions, and goals [15, 79]. Deaf signing individuals often identify themselves as members of a distinct cultural group [108, 112]. Among the most treasured aspects of Deaf culture are sign languages, which function both as a mode of communication and a fundamental component of cultural identity [6, 16, 46]. Despite the historical marginalization of sign languages in education and research, approximately 70 million DHH individuals around the world use sign languages, with over 200 different sign languages in use worldwide [15, 66, 152]. This variability adds to the challenges in creating any sign language technology, in that tools created on the basis of one sign language may not perform well when applied to a different sign language. Across various academic and scientific disciplines, there is a growing consensus that work focusing on sign language is best conducted by groups with linguistic knowledge, alongside authentic cultural knowledge regarding DHH and signing communities [15, 32].

2.2 Sign Language Generation Systems

SLG systems convert written language into signed content. Existing SLG systems typically employ one of two approaches: translating spoken language text directly into pose sequences that represent the corresponding signed translation [69, 70], or incorporating an intermediate written representation between the text and pose sequences [5, 94, 128, 130, 138, 148, 154]. In both cases, the generated

pose sequences are ultimately converted into animations of 3D characters [75, 76] or photorealistic video using generative computer vision models [129, 132, 138].

Research indicates that using an intermediate written representation in SLG systems, preserving linguistic nuances and grammar, results in improved performance [22, 69, 89]. While graphical systems such as SignWriting [139] and HamNoSys [49] offer ways to represent signs, they contain only lexical information and do not contain semantic meaning. Consequently, many SLG systems use sign glosses—a written representation of signs using spoken language text (*e.g.*, English for ASL glosses) that preserves the meaning and grammatical structure of signs [16, 32, 84, 98].

Text-to-gloss translation typically relies on neural machine translation (NMT) models, such as RNNs or Transformers [36, 130, 133, 137, 138, 148, 163], which require extensive labeled data. To address data limitations, some models incorporate syntax-aware adaptations or data augmentation techniques [36, 163]. Nevertheless, alignment with sign language grammar remains a challenge. For example, recent ASL generation systems achieve BLEU-4 scores¹ of less than 0.002 and 0.124 (on a scale from 0 to 1) for translating English text to ASL glosses [71, 163]. Recently, large language models (LLMs) trained on extensive corpora have demonstrated state-of-the-art performance in translation tasks, including for low-resource languages, using few-shot prompting [20, 52, 114], presenting a promising direction for improving SLG systems. In this work, we adopt one of these state-of-the-art LLMs, achieving a BLEU-4 of 0.276, reflecting a compelling translation performance.

The conversion of glosses into pose sequences is generally approached using either motion models that learn sign representations from sub-sequences of motions [130, 131, 154], or from a look-up table that stitch and blend pre-recorded sign sequences [94, 133, 137, 138, 148]. The look-up table approach allows producing full signs based on the dictionary, and the main tasks remain selecting context-appropriate sign variants, and generating smooth and natural sign transitions. Techniques for smoothing transitions include motion graphs, smoothing filters, and frame selection networks [94, 133, 138, 148].

The final step, converting pose sequences into videos, remains an active research area focused on achieving natural, realistic, and temporally consistent results [1, 25, 62, 86, 149, 150]. Early methods used generative adversarial networks (GANs) for motion transfer based on pose data [1, 25, 86, 150]. Following them, SLG works have adapted GANs to generate photorealistic sign videos [132, 148]. Diffusion models have further advanced image and video generation from pose sequences, showing strong results in generating images and videos [42, 62, 63, 96, 123, 126, 161], hence recent SLG work adapted them for generating avatars from pose sequences [39, 40]. However, temporal consistency is not always preserved in these videos, and the animated characters may sometimes cause an uncanny feeling among viewers.

Despite these advancements, challenges remain, particularly in handling non-manual markers (*e.g.*, eyebrow movements) and achieving high-quality outputs with temporal consistency. One

¹BLEU-4 is a machine translation metric representing four gram match between prediction and ground truth. A high BLEU-4 indicates strong alignment with the grammar, where BLEU-4 <20% usually indicate that translations are hard to understand [113].

promising method involves learning a dictionary of facial expressions to match each gloss [148], which enhances visual realism but does not convey additional meaning. This approach often applies the same expression uniformly across sentences, overlooking the contextual nuances of facial expressions and normalizing signers’ faces to face forward, neglecting the subtleties conveyed by directional gaze. Our approach diverges from existing methods by focusing on incorporating non-manual markers while also addressing temporal consistency and enhancing overall visual quality, aiming to create more natural and accurate representation of ASL.

2.3 User Perspectives on Sign Language Technologies

There is growing recognition that developing effective sign language technologies requires a deep understanding of Deaf culture and sign language linguistics, coupled with the refinement of technical approaches and active involvement of the DHH and signing community throughout the design and implementation process [32, 77, 119]. Collaborative and participatory design approaches that incorporate feedback from DHH individuals are essential for creating culturally appropriate and more widely-accepted technologies [16].

Historically, sign language technologies have faced high rejection rates within the Deaf community [47, 61, 147], largely due to top-down design approaches that lack user feedback and a deep understanding of sign languages [77, 93, 118, 160]. For instance, wearable sign language translation gloves have been roundly criticized for focusing narrowly on small sets of handshapes while neglecting other essential linguistic elements like facial expressions and torso orientation [38]. Additionally, such technologies place the communication access burden on Deaf signers rather than hearing individuals, despite being marketed to improve accessibility for the Deaf community [109]. In contrast, sign language technologies developed through active involvement with DHH individuals during the design process have generally been more favorably received [4, 13, 71, 76, 77]. Moreover, DHH users may value technologies that increase their independence and allow for two-way communication, without reliance on cumbersome physical devices [38, 55].

Despite some potential benefits, concerns remain about the accuracy and quality of sign language technologies [35, 65, 76, 77, 80]. A common criticism is that these tools fail to capture the nuances and variations inherent in sign languages, such as personal signing styles and complex grammatical structures, leading to inaccuracies that erode user trust [35, 65, 76, 77, 80]. Historically, technical developments have focused predominantly on single instances of hand shapes while overlooking phrase-level information, facial expressions, and other critical pieces of information [35, 65, 77]. When it comes to SLG tools, such as signing avatars, these limitations are compounded by additional design challenges. User acceptance is influenced by the visual design and movement of signing avatars and the user interface design more generally. Avatars perceived as robotic or as failing to capture the nuances of human signing can hinder effective communication and result in negative user perception [66, 77, 120, 142]. Past work has also recommended reducing the reliance on extensive text-based instructions and offering

customizable features, such as for avatar appearance and signing style [97, 120, 142]. Building on existing literature, this work integrates linguistic and cultural feedback to refine our design choices and improve system performance.

3 Sign Language Generation Prototype

In this section we describe our SLG prototype, which generates ASL videos with manual markers—such as hand shape, location, movements, and palm orientation—as well as non-manual markers, including facial expressions and eyebrow movements. Our focus is on context-free settings, where each sentence is translated independently. We used a modular approach for our system design (Figure 2), allowing increased flexibility and interpretability of each module.

The prototype consists of three components: **Module 1: English Text to ASL Representations**, which leverages a Large Language Model (GPT-4o [2]) to translate an English sentence into English-based ASL glosses and to detect linguistic information relevant to non-manual markers; **Module 2: ASL Representations to Skeletal Pose Sequence**, which takes the LLM outputs and employs a Motion Matching approach to synthesize a skeletal pose sequence; and **Module 3: Skeletal Pose Sequence to ASL Signed Video**, which generates signed video frames representing a photorealistic ASL signer. This modular approach allows for future improvement of the system as the technology advances, by allowing each part to be changed separately. This prototype was iteratively refined within the research team. Insights from these researchers and other collaborators fluent in ASL helped to guide improvements in translation quality, visual and motion quality, and information conveyance.

3.1 Module 1: English Text to ASL Representations

We used an enhanced gloss-based approach that translates an English sentence into an intermediate ASL gloss, including both manual and non-manual information, which is then utilized by subsequent modules. Given the ability of LLMs to naturally absorb and generate grammatical rules, structures, and nuances [20, 122], we used GPT-4o² [2], a state-of-the-art LLM, to perform two key tasks: (1) translate an English sentence into English-based glosses and (2) detect if the English sentence contains linguistic features associated with specific facial expressions (Module 1 in Figure 2). GPT-4o was selected based on our preliminary experiments with various versions of the GPT models. Detailed experimental results are presented in Appendix B.3.

For the first task, we adopted a prompting-based approach using LLMs with “in-context learning” [155], inspired by recent work on low-resource machine translation [48], where dataset sizes are too small to train large-scale translation models. This approach allows the model to adapt and perform specific tasks by interpreting examples or instructions directly embedded in the input text, without requiring explicit retraining [20]. To improve performance, we added 1,494 in-context examples of English sentence-gloss pairs to our prompt from the ASLLRP dataset (representing 80% of the dataset). Given the limited window of GPT-4o (*i.e.*, 128,000 input tokens), which restrict the number of examples that can be included

²Specifically, we used the model gpt-4o-2024-05-13.

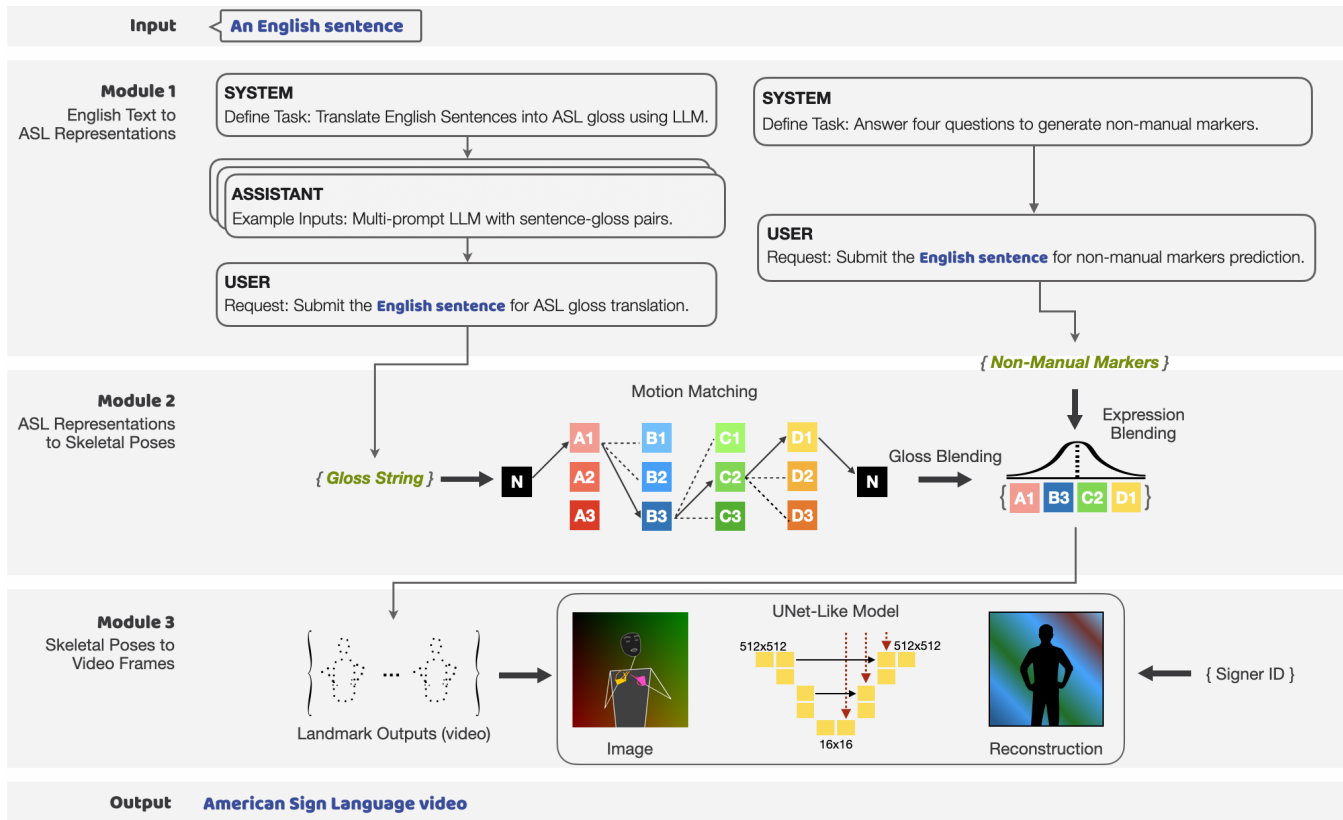


Figure 2: Our prototype includes three self-contained modules. It takes an English sentence as input and generates an ASL video (from top to bottom). Module 1 utilizes a large language model (LLM) to translate the English input into an ASL gloss string and predict non-manual markers. Module 2 employs a Motion Matching approach to generate a skeletal pose sequence from the output of Module 1. Finally, Module 3 uses a UNet-like model, which given an individual signer’s appearance and style (Signer ID), transforms the skeletal pose sequence into signing frames. These are then combined to produce the final signing video.

in a single prompt, we used a “multi-prompting” approach. This method involved splitting the examples into multiple batches and iteratively prompting GPT-4o with each batch. In addition, we asked the LLM to constrain its output by generating glosses within the vocabulary established by our text-to-gloss dictionary described below.

For the second task, we adopted a zero-shot prompting approach, asking the model to predict linguistic features associated with specific facial expressions without any in-context examples. The idea of linguistic predictions was inspired by prior research suggesting that non-manual expressions corresponding to specific grammatical markers, such as raised eyebrows or head tilts, typically involve a consistent set of behaviors that convey meaning within sign language [8, 101]. In this work, we focus primarily on eyebrow movements. To this end, we asked the model to predict whether a given English sentence: is (1) a yes-no question, (2) a wh-question, (3) a conditional statement, and/or (4) contains negation. The outputs from both tasks are then used to generate skeletal poses that are compatible with the subsequent modules, enhancing the integration of non-manual markers.

This approach addresses two common limitations of gloss-based ASL representations: (1) their tendency to deviate from ASL grammar, and (2) their inability to fully capture the context and expressiveness necessary for conveying the full semantics of a sentiment.

Dataset and Implementation Details. After reviewing the available ASL datasets (see Appendix A for more details), we selected the ASLLRP [104] dataset for Module 1. The ASLLRP dataset contains continuous sentence-level ASL videos, isolated ASL videos, ASL glosses, and corresponding English translations. This dataset provides detailed annotations, including textual annotations (e.g., English-based glosses for lexical signs, fingerspelling, classifiers, name signs, and gestures), manual markers (e.g., number of hands used, alternating hand movements), and non-manual markers (e.g., head position and movements, eye gaze, and mouth movements).

Data Preprocessing. While ASLLRP provides the most comprehensive information required for our task, the data is dispersed across various resources and editions. To make effective use of this dataset, we first consolidated these disparate resources into a unified framework, extracting 2,119 English sentence-gloss pairs along with their corresponding signing videos. The signing videos were

then trimmed to isolate specific sign language utterances for our subsequent tasks. To minimize translation errors, we removed gloss annotations that did not alter the overall meaning of the sentence when omitted and standardized all glosses related to fingerspelling. All these changes were done by consulting team members fluent in ASL. We also excluded glosses for classifiers due to their limited sample sizes. After data cleaning, we retained 1,843 English sentence-gloss pairs. Next, we developed a word-gloss dictionary to improve consistency in sign representations of words across different sentences, resulting in 3,915 word-gloss pairs. For the 43 out-of-vocabulary (OOV) words that lacked corresponding videos, we employed fingerspelling as an alternative representation. Finally, four of our researchers conducted a ground truth correction to resolve misalignments between the linguistic labels for the four types of non-manual information and the English text, ensuring the labels more accurately reflected the text content. A more detailed description of our data preprocessing process can be found in Appendix B.1. The conventions used for re-annotating the glosses in this work are summarized in Table D.

LLM Translation and Classification. We used few-shot and zero-shot prompting over GPT-4o [2] to perform these tasks. Our prompts were designed to ensure the outputs could be directly used for downstream tasks and systematic evaluation. Figure 3 overviews the process and prompts, and shows a usage example. More examples of prompts can be found in Table 6. For the translation task, we structured the process by first defining the task for the system. Next, we provided the model with context using English word-gloss pair examples for few-shot learning. Finally, we asked the model to translate each English sentence into ASL glosses, while restricting the translation to our word-gloss dictionary as its vocabulary. For the linguistic features task, we also started by defining the task for the system, and then, using zero-shot prompting, asked the model to classify the linguistic features in the English sentence, *i.e.*, if it contains a yes/no question, a wh-question, a condition, and/or a negation. Zero-shot was enough in this case, because GPT was extensively trained over English text. The exact prompts used for this process are shown in Figure 3.

3.2 Module 2: ASL Representations to Skeletal Pose Sequence

The goal of this module is to take the gloss and non-manual LLM outputs and generate a sequence of skeletal poses at video frame rate, which expresses the input English phrase. We based our approach on Motion Matching, a widely used technique in the Computer Graphics community [21, 27, 58], which takes a large dictionary of short character animations and an input signal and intelligently blends clips together to form a cohesive video. Given the gloss input, a sequence of reference clips is chosen from the dictionary using an optimization function that minimizes the signing concept of “economy of motion.” This principle prioritizes the “best” sign by minimizing the distance between the body position at the end of the previous sign and the start of the next. The selected clips are then linearly blended together to create a cohesive sequence. The non-manual predictions are used as input to an expression blending part of the model which takes the glossed output and augments the facial expressions, in particular targeting eyebrow motion. Our

signing dictionary derived from ASLLRP contains 12,681 signed pose sequences, with many repetitions of each sign, which are labeled with the 3,915 glosses noted above.

Motion Matching typically comprises of three components: (1) a definition for how we represent pose sequences and how they are used for generating the pose sequence dictionary, (2) similarity and optimization functions for identifying the “best” elements for a sequence, and (3) a blending function to create the resulting pose sequence. See Figure 2 (Module 2) for a visual description. The first step chooses and blends the best sign variants. A second step applies expression blending, which augments the pose sequences with non-manual markers to refine facial expressions.

Skeletal Pose Representation & Sign Dictionary. Whole body, face, and hand skeletal keypoints are extracted from all isolated sign videos in ASLLRP using Mediapipe [88], using 3D information for hands and 2D information for the others. We preprocessed this data in three ways. First, we imputed keypoints that were missing due to occlusion issues and poor tracking. For missing keypoints at the beginning or end of a sequence, we filled in points with neutral poses where the hands were positioned together just below the viewpoint from the camera. All other missing keypoints were linearly interpolated using valid keypoints from timesteps before and after within that sequence. One exception was with fingerspelling, where we intentionally kept the non-dominant hand in the same neutral position to avoid jumps between letters in a word. Second, we normalized all keypoints in space so that position and scale of the body and head were consistent across sequences. This alleviated differences in camera position between videos and body shapes between signers. For positioning, we relied on the first frame of each sign with the average shoulder position in subsequent frames relative to that first frame. Lastly, we trimmed the start and end of each sign using annotations from the ASLLRP dataset. For fingerspelling, we sped up the clips to account for the discrepancy between the slower performance in the isolated sign video clips and the faster pace typically used in-situ [121].

Optimization functions. There are many different ways to articulate the same sign for emphasis, style, and convenience [9, 17]. For most signs in our dictionary we have multiple examples of each sign. Often these variants convey the same meaning, but are performed by different signers. Sometimes the meaning does vary. For example, “big” might have versions that convey a medium-big size and a large-big size or a sign might be shown using newer and antiquated styles. In short of having sufficient linguistic information to differentiate sign variations, we select sign variants based on minimizing movement rather than incorporating other linguistic factors. In the signing community this is sometimes referred to as minimizing the “economy of motion,” where an individual may blend together sign variations based on which is physically more efficient. Mathematically, given a vector of keypoint locations $x_{i,t}^p$ where i is a valid gloss index, t is a frame index within a clip, and p is a body part (body, face, hands), we compared the Euclidian distance using a weighted average of the current gloss i and a candidate subsequent gloss indexed by j :

$$d(i, j) = \sum_{p \in \{body, face, hands\}} \alpha_p \cdot \left\| x_{i,T}^p - x_{j,0}^p \right\|_2^2, \quad (1)$$

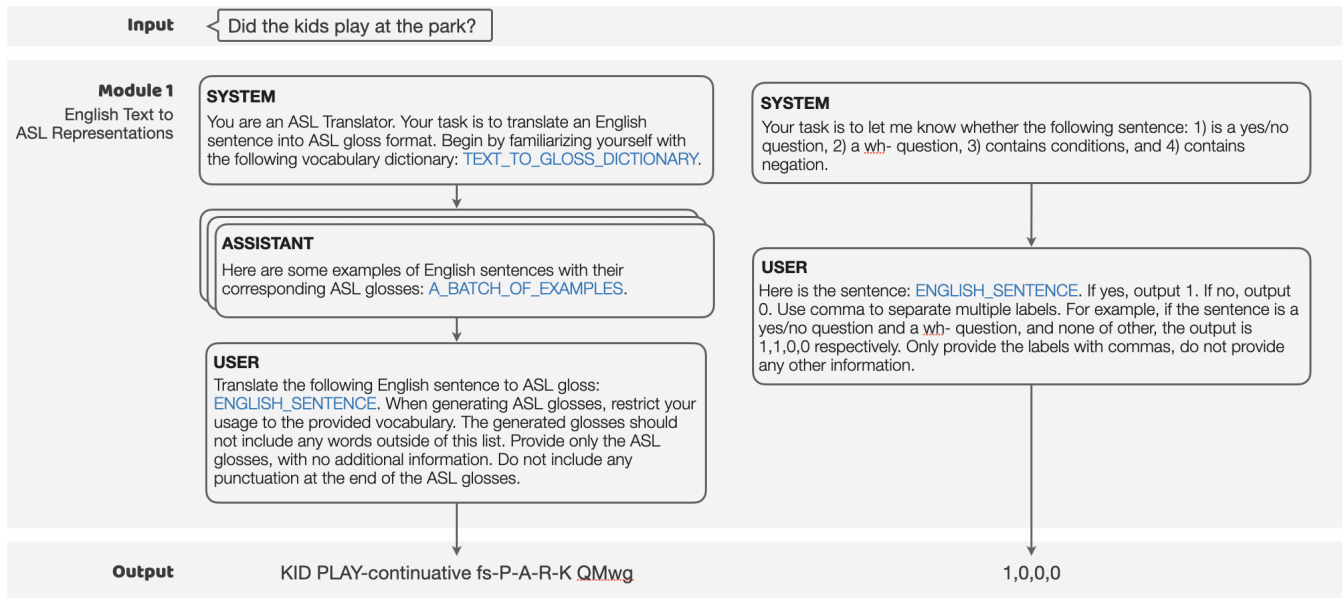


Figure 3: An example from Module 1 showcases two tasks: on the left, translating an English sentence into its corresponding ASL gloss, and on the right, predicting the linguistic features of the same English sentence. `TEXT_TO_GLOSS_DICTIONARY` represents the examples provided to the LLM for each shot. `A_BATCH_OF_EXAMPLES` refers to the examples we provide to the LLM each shot. `ENGLISH_SENTENCE` indicates the user-provided input, which, in this example, “Did the kids play at the park?”

where α_p is a weighting value for each body part, and T is the final frame in the clip. Values of α were chosen to prioritize importance of the body and prevent large changes in posture.

The final sequence of sign videos was determined by minimizing the differences (maximizing the similarity) across all glosses output from the LLM. This was achieved by a greedy algorithm that selected sign videos with the correct gloss labels, prioritizing those where the beginning of the clip was most similar to the end of the previous clip.

Gloss & Expression Blending. We generated a preliminary pose sequence by linearly blending together the start and end of the pose sequence from each chosen gloss instance, using the first and last 20 frames of each clip (at 90 Hz). To increase smooth transitions, we appended half-second neutral pose to the beginning and end of each sequence of videos, which was also interpolated with the gloss videos. We then used the predicted non-manual marker information to augment the facial expressions holistically after stitching the videos together. Specifically, we adjusted the position of the eyebrows throughout the video to reflect whether a sentence was a yes/no question, *wh*-question, or neither. The output of this module is a sequence with body, face, and hand keypoint poses for a full video.

3.3 Module 3: Skeletal Poses to Video Frames

The last module converts the generated pose sequences into a sequence of photorealistic images. First, the input 2/3D skeleton poses are rasterized by drawing the skeletal positions onto an image. Second, these skeletal images are used as input to an image-to-image neural network, which outputs photorealistic images. We choose to

generate videos that resemble “live” signers, in an effort to mitigate confounds that could arise in accurately representing signs with more stylized avatars.

The design decisions regarding the rasterization function—the way the skeleton is drawn—play a critical role in the performance of the image-to-image model. In the baseline rasterization function used by previous work [62, 161], each landmark position was represented by a circle on a 2D image with a monochromatic (black) background, with straight lines connecting the hand and torso landmarks. This is consistent with the commonly used drawing functions within the Mediapipe [88] library. In contrast, in our drawing function, instead of scattered, connected circles, each body part (*i.e.*, hands, body, face) was represented as a convex polygon, with additional connections drawn between the face and the entire body. Each body surface was drawn with a different shade of gray and each hand uses a different color palette where each finger is a different shade. We use the 3D data from each hand to determine the palm orientation (“in” versus “out”) using the surface normal of landmarks surrounding the palm and augment hand colors based on this orientation. Moreover, the background in our dataset varies per person and we find that using a rasterized background color with shades of black-to-red going from top to bottom and black-to-green going from left to right improves the stability of the generated images. The proposed rasterization function significantly improves the image quality and background stability with emphasis to differentiating the hands and individual fingers, disambiguating occlusions originating in overlap between body parts, and differences in the backgrounds of each image.

Although there have been large advances in photorealistic image generation of humans using diffusion models (e.g., ControlNet [161]), results tend to lack temporal consistency and often do not represent hands accurately. Hence, our work builds on image-to-image translation models [19, 53, 73, 143, 161], while adding modern architectures and loss functions. Specifically, we used a U-Net architecture [125], with the encoder and decoder backbones using neural building blocks from the architecture in Imagen [126]. Unlike Imagen, which uses text as an additional input to the system, we condition the decoder using the signers’ identity. This is especially important because we use data from many different Signer IDs as part of the same model. This enables the network to output different visual appearances for each Signer ID in the dataset, which is used when training the network and at inference time.

The model is trained with a combination of three losses. These are an L1 term between the entire generated output frame and the target input frame, an L1 term only on the hand region, and an LPIPS term [162], which is a learned metric that measures perceptual similarity between the output frame and the target input frame. The total loss used to train the model is the sum of the whole frame L1 loss, the hand-specific L1 loss, and the perceptual loss.

Dataset and Implementation Details. Our primary dataset for image generation experiments is How2Sign [34]³. Although it doesn’t contain glosses, in contrast to ASLLRP, which has varied quality across videos, How2Sign contains 35K high-resolution clips of ASL with a vocabulary size of over 16K word tokens. The high resolution and overall data quality of How2Sign helps the model to learn fine-grained and high-quality visual representations of ASL.

While the overall image quality is generally high, there are problems with skeleton tracking, especially when there is significant motion blur or there is ambiguity in hand pose. Thus, when training the model, we discard lower quality frames in efforts to learn more precise mappings between skeletons and photo-realistic humans. We accomplish this by performing automated visual checks in both image and skeletal pose spaces. In image space, we use optical flow to detect motion blur by analyzing the flow vectors between two consecutive frames using Farneback’s method [41]. In pose space, we check for sudden large changes in landmark positions between consecutive frames, which might indicate inaccuracies due to motion blur. Specifically, we compared the current landmark positions to the mean landmark positions over a sliding window of predefined size. While signing, hands tend to move more than the body, so the pose conditions are imposed only for each hand instead of including the entire body and face.

4 Technical Evaluation

We conducted technical evaluations to assess the performance of our proposed system in translating English text into intermediate ASL representations and generating signed videos. The following sections provide a detailed account of each evaluation, including the experimental procedures and the corresponding evaluation results. Note that a direct quantitative comparison with previous

³There is ambiguity as to which individuals in How2Sign gave permission to use their likeness in publications. Thus, for visualization purposes within this paper and supplemental material, we trained additional models that contain the identity of two other people who have given their permission. Qualitatively, these results are representative of the How2Sign results.

work is challenging due to the use of different datasets [71, 95, 163], output modalities, or gloss-less approaches [11]. For instance, while benchmarks for English text-to-ASL gloss translation often use datasets from other languages, benchmarks specific to ASL gloss translation are lacking. Additionally, for video generation, [11] employs the How2Sign dataset and produces SMPL-X 3D human body model poses, whereas our system generates photorealistic videos. These differences in output (3D models vs. photorealistic videos) and their end-to-end design, which precludes comparison of intermediate components, make direct comparisons impractical.

4.1 English Text to ASL Representations

4.1.1 English Text-to-ASL Gloss Translation. We conducted ablation studies to determine the optimal model configuration for translating English sentences into English-based glosses (as illustrated on the left side of Module 1 in Figure 2). Specifically, we examined four key factors: the impact of data preprocessing, the number of in-context examples fed to GPT, the effectiveness of generating glosses within the vocabulary established in our word-to-gloss dictionary, and the necessity of guiding GPT to learn ASL grammar rules⁴. For the number of English-to-gloss examples, we experimented with 600 (33% of dataset) and 1,474 (80% of dataset) sentences from ASLLRP. The dataset was randomly split into a 80/20 ratio to mitigate inconsistencies in distribution. We report BLEU [113] scores (1 to 4 grams) and ROUGE-L [85] scores, two widely used metrics in the machine translation community [11, 40, 128, 130]. Additionally, for a more comprehensive evaluation, we include METEOR [12], CHrF [116], TER [135], and SacreBLEU [117], which are also commonly applied in the literature to assess text-to-gloss translation quality [36, 44, 163].

As shown in Table 1, our ablation study results indicate that data preprocessing improves the LLM’s performance in translating English text to English-based glosses. Similarly, providing the LLM with more examples, when they are chosen randomly, and limiting the generated glosses to those within the word-to-gloss dictionary results in higher BLEU (1 to 4 grams), ROUGE-L, METEOR, and CHrF scores, along with lower TER scores, all of which suggest enhanced model performance.

Interestingly, most experiments showed that adding grammar rules did not improve the model’s translation ability, however, there were some exceptions. For example, when data preprocessing was applied and the LLM was provided with 80% of the entire dataset without limiting the generated glosses to the word-to-gloss vocabulary, we observed mixed results. Specifically, BLEU (1 to 4 grams) scores suggested better model performance without adding grammar rules to the LLM, while other metrics indicated the opposite trend. Furthermore, although direct comparisons are challenging, our system demonstrates compelling translation performance compared to existing results reported in the literature, achieving a BLEU-4 score improvement from 0.191 to 0.276. Table 8 in Appendix B.5 summarizes the existing English Text-to-ASL gloss translation results reported in the literature.

4.1.2 Linguistic Predictions. Falsely predicting linguistic features for a sentence could result in unnecessary non-manual markers

⁴The ASL grammar rules we provided to GPT-4o can be found in B.2 in Appendix.

Table 1: Evaluation results of translating English text to ASL glosses (Task on the left side in Module 1). “Prep.” denotes Preprocessing. *All BLEU-4 and SacreBLEU scores are identical. ↑ indicates that higher values represent better performance, while ↓ indicates that lower values represent better performance. Best results in bold. Note: If “Data Prep.” is set to “No”, the model was not restricted to generating glosses within the word-to-gloss dictionary vocabulary, as the dictionary generation is part of our preprocessing step.

Data Prep.	Number of Examples	Limited Vocab	Grammar Rules	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4* ↑	ROUGE-L ↑	METEOR ↑	CHrF ↑	TER ↓
No	600	-	Yes	0.295	0.204	0.151	0.116	0.573	0.352	0.426	0.668
			No	0.358	0.260	0.201	0.158	0.591	0.386	0.454	0.644
	1,474	-	Yes	0.379	0.280	0.220	0.177	0.603	0.406	0.462	0.625
			No	0.404	0.303	0.239	0.192	0.611	0.432	0.472	0.619
Yes	600 (33% of the entire dataset)	No	Yes	0.470	0.336	0.255	0.197	0.617	0.498	0.487	0.585
			No	0.487	0.355	0.273	0.214	0.627	0.502	0.503	0.572
		Yes	Yes	0.520	0.390	0.305	0.241	0.641	0.530	0.522	0.556
			No	0.520	0.387	0.302	0.237	0.642	0.523	0.528	0.554
	1,474 (80% of the entire dataset)	No	Yes	0.501	0.378	0.298	0.239	0.646	0.534	0.521	0.537
			No	0.513	0.386	0.303	0.243	0.645	0.532	0.519	0.548
		Yes	Yes	0.545	0.415	0.329	0.265	0.662	0.551	0.544	0.524
			No	0.556	0.427	0.341	0.276	0.664	0.560	0.549	0.526

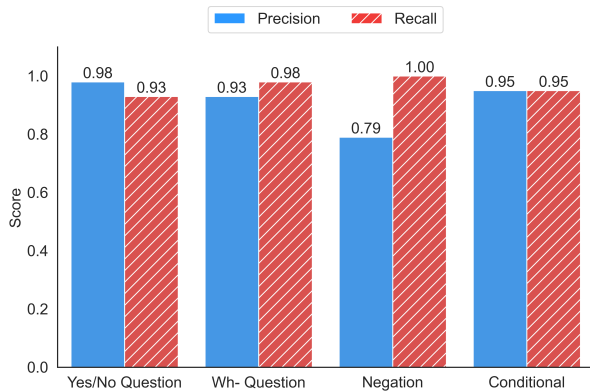


Figure 4: Model performance in detecting linguistic features to generate non-manual marker information. The model demonstrates high performance across all categories, with particularly high recall for detecting negation and precision for detecting yes/no questions. Precision for negation, however, is relatively lower at 0.79.

added to the sequential poses and video frames, potentially leading to confusion in the generated ASL videos. To evaluate the performance of GPT-4o in detecting linguistic features regarding the four questions—whether a sentence is a yes/no question, wh-question, conditional statement, and/or contains negation—we calculate precision and recall for each type of prediction.

Figure 4 summarizes the model’s performance in detecting linguistic features within a given English sentence across the four conditions. Overall, the model demonstrates high accuracy across these tasks, particularly in identifying questions and conditional statements. The relatively low precision for negation (precision=0.79) suggests that the model occasionally incorrectly identified negation in sentences where the human-labeled ground truth did not indicate

negation presence. We analyzed these cases and discovered that in most, the sentences include negative sentiment, e.g., “Why do you hate video games?” or “My sister blamed me but I am innocent!”

4.2 Video Generation

We evaluated our system’s performance in generating signed videos (Modules 2 and 3) using quantitative metrics commonly used for human video generation [149]. These metrics evaluate the generations at either image-level (single-level) or video-level. Image-level metrics include L1, PSNR [60], SSIM [151], LPIPS [162] and FID [54], while video-level metrics include FID-VID [10] and FVD [144]. Following prior research [149], we calculated video-level metrics for sequences of 16 consecutive frames. The dataset contains around 60,000 frames from the How2Sign dataset for training and 15,000 for testing, which correspond to about 40 and 10 minutes of video, respectively. To account for variations in appearance such as clothing, we treated the same signer across different recording sessions as distinct signer identities, resulting in a total of 13 Signer IDs.

We performed several ablation studies to evaluate the efficacy of our design choices. The first ablation study focused on the rasterization function, comparing our proposed enhanced rasterization function with the simpler baseline. The second ablation experiment focused on checking frame quality. Specifically, we reported metrics for our Pose-to-Video model under three conditions: (1) “All frames”, where no frames were excluded from training; (2) “Valid frames”, where frames with missing landmarks were excluded from the training set, and (3) “Proposed”, where frames with missing landmarks, blurry frames, and frames that contain landmarks that indicate temporal inconsistencies were excluded, as detailed in the final paragraph of Section 3.3.

Table 2 presents the evaluation results, demonstrating the proposed approach improves all metrics across the board. The effectiveness of the rasterization function is evident, as the baseline approach produced outputs that resembled a reconstructed skeletal pose rather than a photorealistic human version. The proposed rasterization function provides a better anatomical representation

Table 2: Evaluation results of video generation (Module 3). \uparrow indicates that higher values represent better performance, while \downarrow indicates that lower values represent better performance. Best results in bold.

Experiment	Method	Image				Video		
		L1 \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FID \downarrow	FID-VID \downarrow	FVD \downarrow
Rasterization function	Baseline	31.71E-05	8.069	0.028	1.107	401.14	189.62	2024.42
	Proposed	2.83E-05	23.346	0.864	0.155	56.28	7.23	173.78
Frame quality check	All frames	3.60E-05	19.98	0.838	0.192	187.03	27.71	691.91
	Valid frames	3.14E-05	22.03	0.855	0.165	173.56	15.72	497.17
	Proposed	2.83E-05	23.346	0.864	0.155	56.28	7.23	173.78

of a given pose, enabling the model to learn a more robust mapping between skeletal poses and photorealistic human images. In terms of frame quality, removing lower quality frames progressively improves the model’s performance, reinforcing the conclusion that data quality is just as important as quantity.

5 User Evaluation with DHH Signers

We conducted a user study with 30 DHH participants to further evaluate our prototype system by assessing the perceived quality of our generated signed videos, with a focus on their ASL grammatical correctness—both with and without non-manual markers—understandability, and naturalness of movement. Additionally, we gathered participants’ interest in this technology and its potential use cases. All English sentences were derived from continuous sentence-level signing videos in the ASLLRP dataset. The signed videos presented to participants were either generated from the How2Sign dataset or presented as raw, unprocessed human-signed videos from the ASLLRP dataset.

5.1 Study Design

The survey was conducted online via a web-based survey tool and consisted of two main sections. Participants provided responses through 5-point rating scales and open-ended feedback, allowing for both quantitative and qualitative insights. To minimize bias that might arise from visual aesthetics influencing translation quality evaluations, we intentionally structured the survey to first evaluate visual and motion quality, followed by translation quality. This design choice was inspired by the aesthetic-usability effect, which indicates that users often perceive visually appealing or high-quality visual designs as more functional or accurate [56, 141]. We chose 5-point semantic differential scales, a survey rating scale designed to capture respondents’ attitudes, approaches, and perspectives [110, 111, 159], to gauge DHH participants’ perceptions of the quality of the generated signed videos. A detailed summary of the user study questions is provided in Appendix C.

5.1.1 Section 1: Visual and Motion Quality. This section evaluated Modules 2 (ASL Representations to Skeletal Pose Sequence) and 3 (Skeletal Poses to Video Frames) of our system, focusing on the motion and visual quality of the generated signed videos. The goal was to explore alignments between technical and human evaluations while providing additional assessment of Module 2, which was not fully evaluated during the technical phase due to the lack of established metrics for this module. To achieve this, we presented two types of models for evaluation.

The first type, *AI (Annotations)*, uses human-annotated English-based glosses (manual markers) from the ASLLRP dataset, along with our manually annotated linguistic information (non-manual markers), as input for Modules 2 and 3 of our system. This approach assumes a high-quality text translation and focuses on evaluating the performance of our motion and image models. The second type, *Video Retargeting*, takes skeletal poses extracted from ASLLRP sentence videos as input for Module 3, representing a best-case scenario between these two types. This approach assumes high quality text translation and skeletal extraction, focusing solely on assessing the performance of our visual model and identifying potential issues when retargeting data from the ASLLRP dataset to the How2Sign dataset. Notably, all models using our Module 3 were trained exclusively on the How2Sign dataset (detailed in Section 3.3).

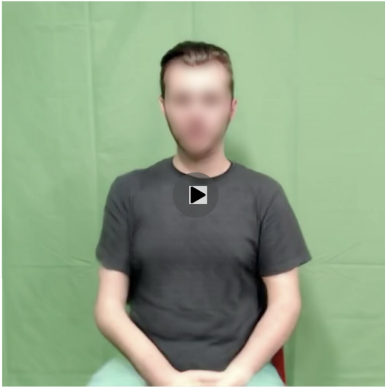
Each participant viewed and rated three videos for each type, where videos were randomly sampled from a larger set of 27 sentences. Participants rated each of the videos on a 5-point rating scale for understandability, visual quality, and naturalness of movement—criteria commonly referenced in the literature [67, 120]. These evaluations were captured across multiple bipolar dimensions, with scale options such as “0 (Very Hard), 1 (Hard), 2 (Neutral), 3 (Easy), 4 (Very Easy)” or “0 (Very Poor), 1 (Poor), 2 (Neutral), 3 (Good), 4 (Excellent).” Note that “N/A” was provided as a default option, but participants were asked to select another response. After completing each rating scale question, participants had the option to provide open-ended feedback for additional insights. Figure 5 provides an example of the survey interface used for this section as presented to the participants.

5.1.2 Section 2: Translation Quality. This section evaluated Module 1 (English text to ASL Representations) of our system, focusing on the translation quality. We aimed to assess how closely our generated ASL aligns with correct ASL grammar and style, and to examine the impact of non-manual markers, specifically facial expressions, on the overall quality and comprehensibility of the signed output. We achieved this by comparing four types of models.

The first type, *AI (Annotations)*, is identical to the approach described in Section 5.1.1. This approach allows us to compare the translation quality of our system with human-annotated ASL representations. The second type, *AI w/o Expr*, uses our full system (Modules 1-3) but with expression blending model turned off to specifically evaluate our system’s effectiveness in generating non-manual markers. The third type, *AI (Full)*, uses our full prototype with the LLM predictions (Modules 1-3), containing both manual and non-manual information. The last type, *Raw Video*, consists of

Watch this video and answer the following questions.

Once you finish, click the next number in the list on the left to move to the next videos.



(a) A video presented to participants.

How easy is it to understand this video?

N/A Very hard Hard Neutral Easy Very Easy

If it is not easy to understand, what could be improved?

Rate the visual quality of the signing in this video. For example, consider facial quality, blurriness.

N/A Very Poor Poor Neutral Good Excellent

If the visual quality is not good or excellent, what could be improved?

Rate how natural the motion is. For example, do the movements and transitions look realistic, and are the signs produced in a typical way?

N/A Very hard Hard Neutral Easy Very Easy

If motion quality is not good or excellent, what could be improved?

(b) Follow-up questions regarding the video.

Figure 5: An example screen from Section 1 of the survey. Videos generated using the How2Sign dataset were presented to participants, followed by a series of evaluation questions. The signer’s face is blurred here to preserve privacy for publication. However, participants viewed an unblurred version during the survey.

original human-signed videos from the ASLLRP dataset without any processing or modeling. The raw videos serve as the best-performing benchmark, providing a reference point for understanding the gap between our system-generated outputs and natural, human-signed videos.

Each participant viewed 21 videos taken from six sentence types. These included a wh-question, a yes/no question, a question that could be mistaken for a statement without non-manual markers, a simpler statement without non-manual markers, a more complex statement involving negation or conditional, and one random sentence with fingerspelling. Within a survey, videos were randomized so that the same sentence was only used once, with one exception. To analyze the expression blending part of our model, we showed each participant the three “question” sentences twice: once using our full system (*AI (Full)*) and once without expression blending (*AI w/o Expr*).

For each video, participants first provided English translations for the ASL content shown. They were then presented with the “true” English translation from the ASLLRP dataset and rated three aspects on a 5-point rating scale: the similarity between the video’s meaning and the “true” English, the quality of the ASL translation (including grammar and signing style), and the accuracy of the facial expressions. To encourage decisive responses and minimize central tendency bias, we adapted scales from prior work [163], excluding the neutral option and using choices such as “0 (Very Poor), 1 (Poor), 2 (Acceptable), 3 (Good), 4 (Excellent).” After completing these ratings, participants used checkbox options and an open-ended text box to report issues with the translations. They also had the options to provide feedback on translation quality and share their ASL interpretation of the English sentence. To maintain consistency and reliability in the evaluation process, each video in both sections was reviewed by at least three participants.

5.1.3 Follow-Up Questions and Demographics. At the end of the survey, participants were asked about their general interest in AI signing technology and its potential use cases. Additionally, demographic information was collected, including gender, age group, the age at which they began learning ASL, their proficiency in both English and ASL, and the frequency of their communication in ASL and spoken English.

5.2 Data Collection

Participants in this study were recruited outside the research group to ensure impartiality and avoid potential biases. To qualify for participation, individuals had to self-identify as DHH, use ASL as their primary language, and be over the age of 18. To ensure participants met these criteria and had the necessary proficiency in ASL, we further implemented a screening process. This process involved prospective participants watching three ASL videos and selecting the corresponding English translations from a set of multiple-choice options. This study went through our organization’s internal user study review process.

After the screening and recruitment process, we enrolled a total of 30 DHH signers who met all eligibility criteria. For demographics, 11 participants were aged 20-29, 10 between 30-39, 7 between 40-49, and 2 between 50-59. Twenty-one participants identified as female, and 9 identified as male. Twenty-four participants learned ASL before age 10, while the remaining learned it later. Regarding proficiency, 23 participants rated their ASL comprehension and production as excellent, while the others rated themselves as good. Fifteen rated their English proficiency as excellent, 11 as good, and 3 as acceptable. All except one reported using ASL daily, with one reporting weekly use. The survey took 45-60 minutes for most individuals to complete.



Figure 6: Descriptive statistics summarize participants’ ratings of motion, visual, and translation quality across model types. Each bar represents the percentage of videos rated within a given response. The right side of each chart (blue) indicates a positive (or neutral) result and the left side (red) indicates a negative result. All models except *Raw video* were trained on the How2Sign dataset to generate signed videos, using English sentences from the ASLLRP dataset as input. In contrast, *Raw video* refers to unprocessed, human-signed videos directly sourced from the ASLLRP dataset.

5.3 Data Analysis

For the rating questions, we report descriptive statistics showing the proportions of each response option for each model type. To account for both fixed and random effects in our data, and to address small sample sizes and deviations from normality in data distributions, we conducted parametric bootstrap linear mixed model (LMM) analyses [30, 115]. These models include model type, sentence type, and participants’ demographic variables—including gender, age category, ASL age, ASL proficiency, and frequency of ASL use—as fixed effects to assess their influence on the ratings. Participant ID was treated as a random effect to capture individual variability. For visual and motion quality evaluations, we conducted three LMM analyses—one each for understanding, visual quality, and naturalness of motion. Similarly, for translation quality evaluations, we conducted another three LMM analyses to assess the similarity of meaning between the generated videos and the English text, the signing quality (focusing on ASL grammar and style), and the accuracy of facial expressions in matching the English text. For open questions, we summarize participants’ feedback to provide insight into their experiences and perceptions.

To further evaluate our system’s translation quality, three authors with ASL experience (1 fluent Deaf signer; 1 fluent hearing signer; 1 novice hearing signer) independently rated the participant-provided translations relative to the English annotations from the ASLLRP dataset. This evaluation assessed whether each translation was semantically equivalent to the target phrase. A 5-point scale was used, defined as follows: 4 = the idea is the same (The same); 3 = the idea is evident but contains one error, such as question changed to a statement, one word error, or one missing element (Similar);

2 = the idea is somewhat similar but unclear or contains multiple errors (Acceptable); 1 = some semblance of the idea is present (Poor); 0 = little to no resemblance to the target (Completely different). Pairwise Pearson correlations [81] were conducted and showed the high agreement among the ratings of the three evaluators, with Pearson’s correlation coefficients ranging from $r = 0.860$ to 0.946 ($p < .001$). For all LMM analyses with model type as a fixed effect, additional pairwise post-hoc comparisons with Holm corrections [59] were conducted to identify specific factors influencing translation quality.

5.4 Findings

5.4.1 Visual and Motion Quality Evaluation Findings. Figure 6a shows results for Section 5.1.1. Regarding the understandability of the generated signed videos from two model types, in the best-case scenario, where raw ASLLRP skeleton data was retargeted using the pose-to-video model from Module 3 (*Video Retargeting*), participants found 60.0% of videos to be *easy* or *very easy* to understand, with 73.3% to be at least *neutral*. Results for *naturalness* were very similar. For visual quality, perceptions were lower, with 32.3% ratings being at least *good* and 60.1% with at least *neutral*. When using our full model with human annotations from the ASLLRP dataset combined with linguistic information from our LLM (*AI (Annotations)*), only 21.1% of ratings indicated the videos were *easy* or *very easy* to understand. Naturalness and visual quality were both rated with lower scores compared to the *Video Retargeting* approach. However, in open-ended responses, some participants commented positively about the body and face movements (e.g., “*Good Body Movements and some lip syncing their words (helpful for*

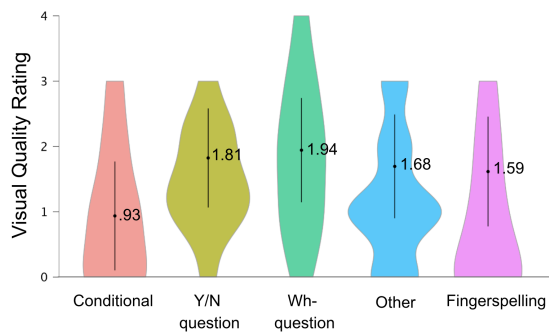


Figure 7: Violin plot illustrating the estimated marginal means for visual quality ratings by sentence type. Wh- and yes-no questions exhibit the highest visual quality ratings, whereas conditional sentences display the lowest ratings (all p values $< .001$). Error bars show 95% confidence intervals.

those who don't understand [the ASL sign]”). Negative sentiments focused on issues like blurriness, cut off fingers, and the need for improved facial expressions. For example, “Blurred background is hard to read the signer” and “[...] fingers cut off sometimes, needs more movement in the facial.”

Our parametric bootstrap LMM analyses revealed significant main effects of model type and sentence type on understandability, visual quality, and naturalness of motion, with few demographic variables showing significant effects. For example, in visual quality ratings, model type showed a significant effect, $\chi^2(1) = 54.53, p < 0.001$, with a bootstrap p -value of 0.002, indicating that the retargeted model received significantly higher visual quality ratings than the *AI (Annotations)* model. Sentence type also had a significant impact on visual quality ratings, $\chi^2(4) = 18.59, p < .001$.

As illustrated in Figure 7, Wh-questions and yes-no questions were rated highest in visual quality, while conditional sentences received the lowest ratings (Holm-corrected post-hoc tests: all $z > 3.7$, all $p < .001$). Among demographic variables, no significant effects were found for gender ($\chi^2(1) = 0.029, p = 0.972$), age ($\chi^2(3) = 1.94, p = 0.584$), ASL age ($\chi^2(2) = 2.54, p = 0.281$), or ASL proficiency ($\chi^2(2) = 2.95, p = 0.229$).

5.4.2 Translation Quality Evaluation Findings. Figure 6b shows results for Section 5.1.2. As expected, the raw videos from the ASLLRP dataset were easiest to understand and had the highest similarity with the English sentences that were shown. Surprisingly, there were a small number of ASLLRP videos that had “poor” or “different” ratings. One participant noted that one of these raw videos had a “Lack of grammar and sentence structure but I can understand what he mean[s].” For the other three models, the differences in translation quality were modest overall—except that the results using our translated glosses (*AI (Full)*) achieved significantly higher ratings than the manually annotated glosses from the ASLLRP dataset (*AI (Annotations)*) in terms of the meaning of the translation. Furthermore, incorporating non-manual markers (i.e., facial expressions) in our full model resulted in higher acceptance compared to the same model without non-manual markers (*AI (w/o Expr)*). The quality of the translation, which focuses on ASL grammar and style, was

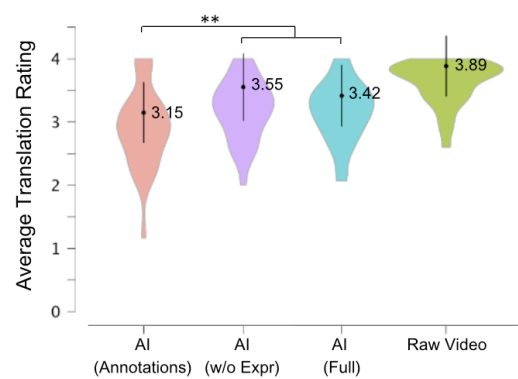


Figure 8: Violin plot illustrating the estimated marginal means for translation quality ratings by model type. *AI (Full)* and *AI (w/o Expressions)* were rated significantly more accurate than *AI (Annotations)*, with no significant difference between them. Error bars represent 95% confidence intervals. While *Raw Video* was rated significantly better than other models, only significance among the other three models is marked for visual simplicity.

rated as at least *acceptable* in 65.3% of cases with our full model. Similarly, the *meaning* of the translation was at least acceptable 53.8% of the time and facial quality was at least acceptable 48.5% of the time.

Similar to the evaluation of visual and motion quality, our LMM analyses revealed a significant main effect of model type on all three aspects of translation quality (all $p < .001$, with bootstrap p -values of 0.002). Contrast analyses showed that the videos generated by the *AI (Annotations)* model were significantly less similar in meaning to the provided English text compared to those produced by our full model (*AI (Full)*; $z = 2.73, p < .05$). However, raw videos consistently received higher ratings than all model-generated outputs. For signing quality and the accuracy of facial expressions, contrast analyses indicated a significant difference between raw videos and all model outputs; however, no significant differences were observed among *AI (Annotations)*, *AI (Full)*, and *AI (w/o Expr)*. Sentence type was also identified as a significant factor influencing translation quality. For example, signing quality ratings exhibited a significant main effect of sentence type ($\chi^2(3) = 227.27, p < .001$, with a bootstrap p -value of 0.002). However, the limited sample size for each sentence type restricted the scope for more detailed analyses. No demographic variables were significant predictors of signing quality ratings.

Our additional LMM analysis, aimed at understanding how well participants' translations aligned with the intended sentences, revealed a strong effect of model type on average translation quality ratings, $\chi^2(3) = 50.45, p < .001$, with a bootstrap p -value of 0.002. As shown in Figure 8, the quality of the translations provided by participants showed significant differences between model types, with *AI (Annotations)* performing significantly worse than both *AI (Full)* and *AI (w/o Expr)* ($z = 3.200$, Holm-corrected $p < .01$). Although participants rated translations with non-manual markers as more acceptable, no significant differences were observed between the

two models using our system, with and without non-manual markers ($z = 0.907$, $p = 0.364$). Sentence type also had a significant effect on the translation quality ratings, $\chi^2(7) = 23.34$, $p < .001$, with a bootstrap p -value of 0.002. In contrast, all demographic variables did not significantly influence translation quality ratings.

Participants reported several issues with both our full model and our model using ASLLRP human gloss annotations, with the majority of concerns pertained to image and motion quality. However, there were a small number of comments on “missing information” and “wrong signs.” One participant noted, “*The signing in the beginning looks very laggy, maybe avoid spelling out the words,*” referring to limitations in fingerspelling where individual letters appeared to jump between locations or were signed slowly compared to natural signing.

5.4.3 Interest of AI Signing Technology and Its Use Cases. Participants expressed varying levels of interest in AI signing technology. One highly enthusiastic participant remarked, “*Everything looks good so far, most of the ASL is correct, definitely on the right path. This would be a great tool and technology for those who struggle with communication in the hearing community. It’s super convenient and I can’t wait. Thank you for allowing me to be a part of this,*” indicating their inclination to sign up to use such a technology in the future. The least interested person highlighted that the quality of the technology is far from being useful, stating, “*I am not interested seeing AI signing technology because it’s too complicated to understand the ASL signer.*” Despite this, the same participant later expressed that the technology could be valuable for certain use cases.

When asked about their interest in photorealistic, cartoon, or 3D avatars to represent AI signers, participants provided mixed feedback, but with a lean towards photorealistic styles. One participant emphasized the value of realism, stating, “*Realistic and Authentic[—]it is simpler for viewers to relate to and believe in the content when a live signer offers an honest and realistic experience. It better for training and teaching other people ASL.*” Stylized signers could be of interest for social media, advertisements, or children’s content, but multiple people noted the importance of ensuring the stylized depiction is capable of conveying nuance of sign language: “*I think more stylized appearance can do, but needs [to be] clear in image and facial expressions.*”

Participants mentioned a wide range of potential use cases for AI signing technology. Many of these examples related to simultaneous recognition and generation of ASL for real-time social interactions. Others focused on one-sided interactions, such as ASL generation of live presentations.

6 Discussion

Our goal was to develop a prototype ASL generation system, addressing key challenges limiting real-world applicability of existing SLG systems, and to explore whether DHH signers would find this technology useful. Below, we reflect on our design process, provide key insights learned, identify areas for improvement, and discuss computational and ethical considerations in the use of our system.

6.1 Technical Insights from the Design and Evaluation Process

During the design process and evaluations with DHH participants, we gained valuable technical insights that informed our choices and identified areas for future improvement. One key finding was the importance of careful data handling for translation tasks. Our ablation study results, as shown in Table 1, highlight the importance of data preprocessing, increasing the number of examples provided to the model, and constraining the translation within the pre-generated vocabulary to improve model translation performance in the low-resource settings. Considerable effort was dedicated to creating an annotation scheme that not only accurately represents ASL signs and sentences but also functions effectively when used with the LLM and the rest of our prototype. This points to a fundamental challenge with glossing: the diverse definitions and interpretations of ASL glosses. Standardization across datasets could mitigate this issue and improve accuracy by allowing the combination of different data resources [16].

Our use of an LLM for generating both manual and non-manual information demonstrate potential, with the model achieving a BLEU-4 of 0.276 for translating English sentences from the ASLLRP dataset into ASL glosses. While direct comparisons—such as running our dataset on other systems or applying our system to other datasets—are challenging due to the inaccessibility of other datasets and systems, this represents highest reported score for such translation task in the literature, highlighting the effectiveness of few-shot prompting techniques in handling low-resource languages. More than half of the time, DHH participants found the meaning of the generated videos “Acceptable”, “Similar” or “The Same” when compared to the English text. However, in close to 50% of the examples, they rated our translations as “Poor” or “Very poor” concerning ASL grammar and style, indicating a need for further improvement in aligning the output with native signing conventions.

Our additional experiments on English Text-to-ASL gloss translation using Retrieval Augmented Generation (RAG) [82] demonstrated improved performance, achieving a BLEU-4 score of 0.279 ± 0.003 . These results suggest potential for further enhancement in translation accuracy. Detailed descriptions of the experiments are provided in Appendix B.4. Beyond translation accuracy, our innovation on extracting non-manual markers directly from the English text using zero-shot prompting, could potentially enhance the naturalness and grammatical accuracy of the generated videos. Nonetheless, some linguistic features were misidentified due to inconsistencies between gloss annotations and English sentences (as discussed in Section 4.1.2), suggesting the need for prompt fine-tuning or more targeted examples.

The use of a Motion Matching approach for generating skeletal pose sequences offered both promise and challenges. By optimizing for “economy of motion,” this method enabled smoother transitions between signs, contributing to more fluid and natural signing overall. However, we encountered issues with fingerspelling, where unintended movements appeared between letters, disrupting the continuity of motion. This challenge was also noted in user feedback, highlighting gaps in achieving the desired naturalness in coarticulations, particularly for complex cases such as fingerspelling. The noticeable naturalness rating difference between the

full model and the retargeted approach—where only in 34.5% of the cases participants perceived the naturalness of our videos as “Neutral” or better, compared to 71.1% for the retargeted version (results shown in Figure 6a)—emphasizes the need for refining our skeletal motion generation method.

A key factor limiting the adoption of existing SLG systems by DHH users is the low quality of the generated signing videos, which are often described as robotic or blurry [66, 77, 120, 142]. Our technical evaluations, as detailed in Table 2, demonstrate that our approach improves the visual quality of the generated videos by systematically eliminating data errors, such as missing landmarks, blurriness, and temporal inconsistencies in landmark positioning, through using only the highest-quality frames. However, we still observe a gap between these technical improvements and practical usability, as in 77.8% of the time DHH participants found the visual quality of our signing videos to be “Poor” or “Very Poor”. Additionally, participants noted that head movements did not consistently align with the camera. Future work could explore integrating more advanced generative models such as diffusion models [29, 156] to enhance video quality.

6.2 A Need for Larger, High-quality, and Comprehensive ASL Datasets

Despite using the largest and highest-quality ASL datasets available, the chosen datasets still suffer from several limitations. The ASLLRP dataset is advantageous in that it contains tens of thousands of videos with comprehensive annotations (e.g., glosses, English sentences, non-manual markers). However, the dataset suffers from limited visual quality due to issues such as image resolution and motion blur, which proved challenging for generating compelling image-to-image models during our initial experiments. When we turned to the How2Sign dataset for training image-to-image models, we found that the visual quality was significantly better. However, this introduced inconsistencies between datasets. For example, signers in the ASLLRP dataset tend to be seated and looking at prompts away from the camera; while signers in the How2Sign dataset maintain direct eye contact and are looking closer at the camera, slightly sideways. These discrepancies, coupled with the relatively small size of these datasets, highlight the need for more comprehensive and consistent ASL datasets.

Furthermore, the reliance on human-labeled gloss annotations in existing ASL datasets introduces multiple sources of errors and inconsistencies. While many English sentences in the ASLLRP dataset are derived from context-free ASL utterances translated into glosses and English text, others come from longer narrative videos. In these cases, accurate translation requires full contextual understanding, which the annotations may not always provide [140]. Consequently, for many of these context-dependent sentences, our text-to-gloss translations may be more accurate than the original human annotated glosses. This is reflected in our model’s performance, where we achieved a BLEU-4 score of 0.305 without these context-dependent sentences (52 in the test set), compared to 0.276 with them. Further supporting this observation, our user study indicated that DHH users rated the quality of our translations, specifically regarding the meaning of video compared to the English text, to be more acceptable than the manually annotated glosses provided by

the ASLLRP dataset. These findings highlight the critical need for more robust, high-quality datasets with standardized annotation practices to support the development of effective SLG systems [16].

6.3 Addressing the Complexities of ASL in Sign Language Generation Technologies

The complexities of ASL grammar present challenges for developing effective SLG technologies. While general guidelines for ASL grammar exist, the language, like all natural languages, does not always adhere to rigid grammatical structures in everyday use. This complexity is evident in the mixed results from our experiments, where attempts to provide grammar guidelines to the LLM did not consistently enhance translation performance. Many examples in the ASLLRP dataset, while grammatically correct, diverge from these general guidelines (as shown in Table 1). Feedback from our user study, which highlighted stylistic and grammatical errors, emphasizes the need for a more nuanced computational understanding of how ASL is used in diverse, real-world contexts to improve. Furthermore, regional variations within a single language and differences across multiple sign languages introduce additional layers of complexity that remain to be addressed.

This work studies several aspects of both manual and non-manual markers in ASL morphology, lexicon, and syntax, such as compounds, agreement verbs (directional verbs indicating agreement with the subject and object), fingerspelling, and name signs (more details can be found in Table D). However, these linguistic features are analyzed only within the context of the dataset used in this study, which does not capture the full range of their usage in ASL. Additionally, several other facets of ASL grammar and usage remain unexplored. For example, we excluded one type of manual marker, classifiers, due to limited data available to model them accurately. Classifiers, which are essential for conveying nuanced meanings and spatial relationships in ASL, require context-aware data and more sophisticated modeling approaches. As SLG systems evolve towards context-dependent applications, incorporating classifiers will be critical for enhancing the naturalness and expressiveness of the generated signs. Additionally, our work focuses primarily on eyebrow movements, one type of facial expressions within non-manual markers, used to indicate questions, conditional statements, and negation. However, non-manual markers in ASL consist of a wide range of features, including head tilts, mouth shapes, and body posture, which also contribute to the grammar and meaning of signed sentences [18, 78, 136]. Future work is needed to expand the modeling of these additional markers to capture the full complexity of ASL.

Moreover, our study focused on context-free SLG, where each sentence is generated independently. However, sign languages heavily use indexing and spatial referencing, such as referencing people or places mentioned earlier in a conversation [45, 153]. Our current prototype system lacks the capacity to remember or track these spatial references over multiple utterances. Additionally, types of signing like storytelling often involve more extensive use of expressions, classifiers, spatial references, and role shifting than our prototype can currently support. Addressing these challenges will require more data, modeling, and interdisciplinary collaboration with ongoing feedback from the DHH and signing communities.

6.4 Computational and Ethical Considerations

While both technical and human evaluations demonstrate the potential of our prototype system, and the modular approach offers flexibility by enabling individual components to be improved or replaced as technologies advance, there are several computational and ethical considerations that should be carefully addressed when using or further improving the system. First, the current prototype requires running GPT-4o inference for every generation instance with longer prompts, which introduces computational and financial costs, as well as scalability challenges, particularly for real-time or large-scale applications. Optimization techniques or lighter models may need to be explored to address this issue. Second, the nature of modular approach can lead to the loss of information between stages, computational inefficiencies, or biases imposed by external constraints at each module. Addressing these shortcomings will require careful integration of modules. Third, the use of LLMs might pose a risk of generating inappropriate or offensive language, which could introduce harm to the DHH community or undermine their trust in using such system. As emphasized in both academia and industry (e.g., Apple’s Responsible AI white paper [72]), designing AI tools with care to proactively mitigate potential harms must be a top priority. This includes implementing content filtering mechanisms, rigorous validation processes, and culturally sensitive design practices to ensure that the system outputs are respectful, inclusive, and aligned with community expectations.

7 Conclusion

In this paper, we proposed a prototype ASL generation system aimed at improving the naturalness, comprehensiveness, and overall quality of generated signs, addressing key limitations in existing approaches. Our technical evaluations indicate that our proposed approaches improve these aspects, enhancing the quality of generated ASL content. Feedback from DHH participants was mixed; while there was general interest in the system, concerns regarding visual quality and naturalness were noted. Reflecting on our design process and study findings, we discuss key insights and identify key areas for future improvement. While further work is needed, our study takes an initial step toward developing sign language generation systems that better meet the needs of the DHH and signing communities, offering real-world value.

Acknowledgments

We sincerely thank all reviewers for their valuable feedback, which significantly enhanced our work. We also extend our gratitude to the participants of our user study for their time and contributions. Lastly, we deeply appreciate Gus Shitama, Julia Sohnen, Pooja Solanki, Sheridan Laine, and Antony Kennedy for their insightful discussions and support with study-related tasks.

References

- [1] K Aberman, M Shi, J Liao, D Lischinski, B Chen, and D Cohen-Or. 2019. Deep Video-Based Performance Cloning. In *Computer Graphics Forum*, Vol. 38. Wiley-Blackwell Publishing Ltd., 219–233.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [3] Mohamed Amin, Hesahm Hefny, and Ammar Mohammed. 2021. Sign Language Gloss Translation using Deep Learning Models. *International Journal of Advanced Computer Science and Applications* 12 (Jan. 2021). <https://doi.org/10.14569/IJACSA.2021.0121178>
- [4] Vidia Anindhita and Dessi Puji Lestari. 2016. Designing interaction for deaf youths by using user-centered design approach. In *2016 international conference on advanced informatics: Concepts, theory and application (icaicta)*. IEEE, 1–6.
- [5] Rotem Shalev Arkushin, Amit Moryossef, and Ohad Fried. 2023. Ham2pose: Animating sign language notation into pose sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21046–21056.
- [6] British Deaf Association. Veditz Quote - 1913 (2015). <https://vimeo.com/132549587>
- [7] Charlotte Baker-Shenk. 1985. The facial behavior of deaf signers: Evidence of a complex language. *American Annals of the Deaf* 130, 4 (1985), 297–304.
- [8] Charlotte Lee Baker-Shenk. 1983. *A microanalysis of the nonmanual components of questions in American Sign Language*. University of California, Berkeley.
- [9] Charlotte Lee Baker-Shenk and Dennis Cokely. 1991. *American Sign Language: A teacher’s resource text on grammar and culture*. Gallaudet University Press.
- [10] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. 2019. Conditional GAN with Discriminative Filter Generation for Text-to-Video Synthesis. In *IJCAI*, Vol. 1. 2.
- [11] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. 2024. Neural Sign Actors: A diffusion model for 3D sign language production from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1985–1995.
- [12] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (Eds.). Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. <https://aclanthology.org/W05-0909>
- [13] Patrick Boudreault, Muhammad Abubakar, Andrew Duran, Bridget Lam, Zehui Liu, Christian Vogler, and Raja Kushalnagar. 2024. Closed Sign Language Interpreting: A Usability Study. In *International Conference on Computers Helping People with Special Needs*. Springer, 42–49.
- [14] Danielle Bragg, Naomi Caselli, John W Gallagher, Miriam Goldberg, Courtney J Oka, and William Thies. 2021. ASL sea battle: gamifying sign language data collection. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.
- [15] Danielle Bragg, Naomi Caselli, Julie A Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E Ladner. 2021. The fate landscape of sign language ai datasets: An interdisciplinary perspective. *ACM Transactions on Accessible Computing (TACCESS)* 14, 2 (2021), 1–45.
- [16] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Pittsburgh PA USA, 16–31. <https://doi.org/10.1145/3308561.3353774>
- [17] DIANE Brentari. 1998. A prosodic model of sign language phonology. *A Bradford Book* (1998).
- [18] Diane Brentari and Laurinda Crossley. 2002. Prosody on the hands and face: Evidence from American Sign Language. *Sign Language & Linguistics* 5, 2 (2002), 105–130.
- [19] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.
- [20] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165* (2020).
- [21] Michael Büttner and Simon Clavet. 2015. Motion matching—the road to next gen animation. *Proc. of Nucl. ai 1*, 2015 (2015), 2.
- [22] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, 7784–7793. <https://doi.org/10.1109/CVPR.2018.00812>
- [23] Brenda Cartwright. 2024. Signing Savvy. <https://www.signingsavvy.com/index.php>
- [24] Naomi K Caselli, Zed Sevcikova Sehyr, Ariel M Cohen-Goldberg, and Karen Emmorey. 2017. ASL-LEX: A lexical database of American Sign Language. *Behavior research methods* 49 (2017), 784–801.
- [25] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5933–5942.
- [26] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems* 35 (2022), 17043–17056.

- [27] Simon Clavet et al. 2016. Motion matching and the road to next-gen animation. In *Proc. of GDC*, Vol. 2. 4.
- [28] Onno Crasborn and Han Sloetjes. 2008. Enhanced ELAN functionality for sign language corpora. In *6th International Conference on Language Resources and Evaluation (LREC 2008)/3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*. 39–43.
- [29] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 10850–10869.
- [30] Anthony Christopher Davison and David Victor Hinkley. 1997. *Bootstrap methods and their application*. Number 1. Cambridge university press.
- [31] Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumpfrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2024. ASL citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems* 36 (2024).
- [32] 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. *arXiv preprint arXiv:2403.02563* (2024).
- [33] Philippe Dreuw, Carol Neidle, Vassilis Athitsos, Stanley Sclaroff, and Hermann Ney. 2008. Benchmark databases for video-based automatic sign language recognition. In *Proceedings of Language Resources and Evaluation Conference (LREC) 2008*. EUROPEAN LANGUAGE RESOURCES ASSOC-ELRA.
- [34] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metzger, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Nashville, TN, USA, 2734–2743. <https://doi.org/10.1109/CVPR46437.2021.00276>
- [35] Sarah Ebling and John Glauert. 2016. Building a Swiss German Sign Language avatar with JASigning and evaluating it among the Deaf community. *Universal Access in the Information Society* 15 (2016), 577–587.
- [36] Santiago Egea Gómez, Euan McGill, and Horacio Saggion. 2021. Syntax-aware Transformers for Neural Machine Translation: The Case of Text to Sign Gloss Translation. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, Reinhard Rapp, Serge Sharoff, and Pierre Zweigenbaum (Eds.). INCOMA Ltd., Online (Virtual Mode), 18–27. <https://aclanthology.org/2021.bucc-1.4>
- [37] Karen Emmorey. 2001. *Language, cognition, and the brain: Insights from sign language research*. Psychology Press.
- [38] Michael Erard. 2017. Why Sign-Language Gloves Don't Help Deaf People. <https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/>
- [39] Sen Fang, Chunyu Sui, Xuedong Zhang, and Yapeng Tian. 2023. SignDiff: Learning Diffusion Models for American Sign Language Production. *arXiv preprint arXiv:2308.16082* (2023).
- [40] Sen Fang, Lei Wang, Ce Zheng, Yapeng Tian, and Chen Chen. 2024. Sign-LLM: Sign Languages Production Large Language Models. *arXiv preprint arXiv:2405.10718* (2024).
- [41] Gunnar Farneback. 2003. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*. Springer, 363–370.
- [42] Mengyang Feng, Jinlin Liu, Kai Yu, Yuan Yao, Zheng Hui, Xiefan Guo, Xianhui Lin, Haolan Xue, Chen Shi, Xiaowen Li, et al. 2023. Dreamoving: A human video generation framework based on diffusion models. *arXiv e-prints* (2023), arXiv–2312.
- [43] The Academic Center for Excellence. 2023. ASL Grammar Guide. <https://germana.edu/sites/default/files/2023-07/ASL%20Grammar%20Guide%20%28edit%207-24-23%29.pdf>
- [44] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. [n. d.]. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. ([n. d.]).
- [45] Lynn A Friedman. 1975. Space, time, and person reference in American Sign Language. *Language* (1975), 940–961.
- [46] Neil Stephen Glickman. 1993. *Deaf identity development: Construction and validation of a theoretical model*. University of Massachusetts Amherst.
- [47] Jan Gugenheimer, Katrin Plaumann, Florian Schaub, Patrizia Di Campli San Vito, Saskia Duck, Melanie Rabus, and Enrico Rukzio. 2017. The impact of assistive technology on communication quality between deaf and hearing individuals. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 669–682.
- [48] Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and He-Yan Huang. 2024. Teaching Large Language Models to Translate on Low-resource Languages with Textbook Prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 15685–15697.
- [49] Thomas Hanke. 2004. HamNoSys-representing sign language data in language resources and language processing contexts. In *LREC*, Vol. 4. 1–6.
- [50] Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. Extending the Public DGS Corpus in size and depth. In *Proceedings of the LREC2020 9th workshop on the representation and processing of sign languages: Sign language resources in the service of the language community, technological challenges and application perspectives*. 75–82.
- [51] Vicki L Hanson and Carol A Padden. 2012. Computers and videodisc technology for bilingual ASL/English instruction of deaf children. In *Cognition, Education, and Multimedia*. Routledge, 49–63.
- [52] Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210* (2023).
- [53] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).
- [54] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).
- [55] Joseph Hill. 2020. Do deaf communities actually want sign language gloves? *Nature Electronics* 3, 9 (2020), 512–513.
- [56] JoAndrea Hoegg, Joseph W Alba, and Darren W Dahl. 2010. The good, the bad, and the ugly: Influence of aesthetics on product feature judgments. *Journal of Consumer Psychology* 20, 4 (2010), 419–430.
- [57] Annette Hohenberger, Daniela Happ, and Helen Leuninger. 2002. Modality-dependent aspects of sign language production: Evidence from slips of the hands and their repairs in German Sign Language. *Modality and structure in signed and spoken languages* (2002), 112–142.
- [58] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. 2020. Learned motion matching. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 53–1.
- [59] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.
- [60] Alain Hore and Djemel Ziou. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*. IEEE, 2366–2369.
- [61] Jeremy Hsu. 2024. AI can turn text into sign language – but it's often unintelligible. <https://www.newscientist.com/article/2436111-ai-can-turn-text-into-sign-language-but-its-often-unintelligible>
- [62] Li Hu. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8153–8163.
- [63] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. Composer: creative and controllable image synthesis with composable conditions. In *Proceedings of the 40th International Conference on Machine Learning*. 13753–13773.
- [64] Matt Huenerfauth. 2008. Generating American Sign Language animation: overcoming misconceptions and technical challenges. *Universal Access in the Information Society* 6 (2008), 419–434.
- [65] Matt Huenerfauth. 2009. A linguistically motivated model for speed and pausing in animations of american sign language. *ACM Transactions on Accessible Computing (TACCESS)* 2, 2 (2009), 1–31.
- [66] Matt Huenerfauth and Vicki Hanson. 2009. Sign language in the interface: access for deaf signers. *Universal Access Handbook*. NJ: Erlbaum 38 (2009), 14.
- [67] Matt Huenerfauth, Liming Zhao, Erdan Gu, and Jan Allbeck. 2007. Evaluating American Sign Language generation through the participation of native ASL signers. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*. 211–218.
- [68] Matt Huenerfauth, Liming Zhao, Erdan Gu, and Jan Allbeck. 2008. Evaluation of American Sign Language Generation by Native ASL Signers. *ACM Transactions on Accessible Computing* 1, 1 (May 2008), 1–27. <https://doi.org/10.1145/1361203.1361206>
- [69] Eui Jun Hwang, Sukmin Cho, Huije Lee, Youngwoo Yoon, and Jong C Park. 2024. Universal Gloss-level Representation for Gloss-free Sign Language Translation and Production. *arXiv preprint arXiv:2407.02854* (2024).
- [70] Eui Jun Hwang, Huije Lee, and Jong C Park. 2024. A Gloss-Free Sign Language Production with Discrete Representation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–6.
- [71] Mert Inan, Katherine Atwell, Anthony Sicilia, Lorna Quandt, and Malihe Alikhani. 2024. Generating Signed Language Instructions in Large-Scale Dialogue Systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*. 140–154.
- [72] Apple Inc. 2024. Introducing Apple's On-Device and Server Foundation Models. <https://machinelearning.apple.com/research/introducing-apple-foundation-models>
- [73] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

- [74] Hamid Reza Vaezi Joze and Oscar Koller. 2018. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053* (2018).
- [75] Jung-Ho Kim, Eui Jun Hwang, Sukmin Cho, Du Hui Lee, and Jong C Park. 2022. Sign language production with avatar layering: A critical use case over rare words. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 1519–1528.
- [76] Michael Kipp, Alexis Heloir, and Quan Nguyen. 2011. Sign language avatars: Animation and comprehensibility. In *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15–17, 2011. Proceedings 11*. Springer, 113–126.
- [77] Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. 2011. Assessing the deaf user perspective on sign language avatars. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*. 107–114.
- [78] Edward S Klima and Ursula Bellugi. 1979. *The signs of language*. Harvard University Press.
- [79] Paddy Ladd. 2003. *Understanding deaf culture: In search of deafhood*. Multilingual Matters.
- [80] Sooyeon Lee, Abraham Glasser, Becca Dingman, Zhaoyang Xia, Dimitris Metaxas, Carol Neidle, and Matt Huenerfauth. 2021. American sign language video anonymization to support online participation of deaf and hard of hearing users. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–13.
- [81] Joseph Lee Rodgers and W Alan Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician* 42, 1 (1988), 59–66.
- [82] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [83] Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. 2020. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems* 33 (2020), 12034–12045.
- [84] Scott K Liddell. 2003. *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press.
- [85] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. <https://aclanthology.org/W04-1013>
- [86] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeonwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. 2019. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)* 38, 5 (2019), 1–14.
- [87] Ceil Lucas. 2001. *The sociolinguistics of sign languages*. Cambridge University Press.
- [88] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).
- [89] Xiaohan Ma, Rize Jin, and Tae-Sun Chung. 2024. Multi-Channel Spatio-Temporal Transformer for Sign Language Production. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 11699–11712.
- [90] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. 2018. SignFi: Sign language recognition using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–21.
- [91] Aleix M Martínez, Ronnie B Wilbur, Robin Shay, and Avinash C Kak. 2002. Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE, 167–172.
- [92] Ross E Mitchell and Travas A Young. 2023. How many people use sign language? A national health survey-based estimate. *Journal of Deaf Studies and Deaf Education* 28, 1 (2023), 1–6.
- [93] David C Mohr, Ken R Weingardt, Madhu Reddy, and Stephen M Schueller. 2017. Three problems with current digital mental health research... and three things we can do about them. *Psychiatric services* 68, 5 (2017), 427–429.
- [94] Amit Moryossef, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling. 2023. An open-source gloss-based baseline for spoken to signed language translation. In *Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages*. 22–33.
- [95] Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data Augmentation for Sign Language Gloss Translation. <http://arxiv.org/abs/2105.07476> arXiv:2105.07476 [cs].
- [96] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4296–4304.
- [97] Laura J Muir and Iain EG Richardson. 2005. Perception of sign language and its application to visual communications for deaf people. *Journal of Deaf studies and Deaf education* 10, 4 (2005), 390–401.
- [98] 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 682–693. <https://doi.org/10.18653/v1/2023.acl-short.60>
- [99] Jemina Napier. 2002. *Sign language interpreting: Linguistic coping strategies*. Douglas McLean.
- [100] Carol Neidle. 2001. SignStream™: A database tool for research on visual-gestural language. *Sign language & linguistics* 4, 1-2 (2001), 203–214.
- [101] Carol Neidle. 2002. Signstream annotation: Addendum to conventions used for the american sign language linguistic research project, Report No. 11. (2002).
- [102] Carol Neidle. 2007. Signstream annotation: Addendum to conventions used for the american sign language linguistic research project. (2007).
- [103] Carol Neidle. 2017. A User’s guide to SignStream® 3. (2017).
- [104] Carol Neidle, Augustine Opoku, and Dimitris Metaxas. 2022. ASL Video Corpora & Sign Bank: Resources Available through the American Sign Language Linguistic Research Project (ASLLRP). <http://arxiv.org/abs/2201.07899> arXiv:2201.07899 [cs].
- [105] Carol Neidle, Augustine Opoku, and Dimitris Metaxas. 2022. Asl video corpora & sign bank: Resources available through the american sign language linguistic research project (asllrp). *arXiv preprint arXiv:2201.07899* (2022).
- [106] Carol Neidle and Christian Vogler. 2012. A new web interface to facilitate access to corpora: Development of the ASLLRP data access interface (DAD). In *Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC*, Vol. 3. Citeseer, 23–28.
- [107] Form Sign Datasets Carol Neidle and Augustine Opoku. [n. d.]. *Boston University, Boston*. Technical Report. MA Report.
- [108] Chijioke Obasi. 2008. Seeing the deaf in “deafness”. *Journal of Deaf Studies and Deaf Education* 13, 4 (2008), 455–465.
- [109] Visual Anthropology of Japan. 2019. “Why Sign-Language Gloves Don’t Help Deaf People” -and- neither does the “Woman’s hand” iPhone case to keep you company” -and then- a couple of new products that were made with deaf collaboration. <http://visualanthropologyofjapan.blogspot.com/2019/07/why-sign-language-gloves-dont-help-deaf.html>
- [110] Charles E Osgood. 1957. The measurement of meaning. *Urbana: University of Illinois Press* (1957).
- [111] Charles E Osgood. 1964. Semantic differential technique in the comparative study of cultures. *American anthropologist* 66, 3 (1964), 171–200.
- [112] Carol A Padden and Tom L Humphries. 1988. *Deaf in America: Voices from a culture*. Harvard University Press.
- [113] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. <https://doi.org/10.3115/1073083.1073135>
- [114] Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards Making the Most of ChatGPT for Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 5622–5633.
- [115] José Pinheiro and Douglas Bates. 2006. *Mixed-effects models in S and S-PLUS*. Springer science & business media.
- [116] Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina (Eds.). Association for Computational Linguistics, Lisbon, Portugal, 392–395. <https://doi.org/10.18653/v1/W15-3049>
- [117] Matt Post. 2018. A call for clarity in reporting BLEU scores. *arXiv preprint arXiv:1804.08771* (2018).
- [118] Soraia Prieth, J Alfredo Sánchez, and Josefina Guerrero. 2022. A systematic review of user studies as a basis for the design of systems for automatic sign language processing. *ACM Transactions on Accessible Computing* 15, 4 (2022), 1–33.
- [119] Soraia Prieth, J. Alfredo Sánchez, and Josefina Guerrero. 2022. A Systematic Review of User Studies as a Basis for the Design of Systems for Automatic Sign Language Processing. *ACM Transactions on Accessible Computing* 15, 4 (Dec. 2022), 1–33. <https://doi.org/10.1145/3563395>
- [120] Lorna C Quandt, Athena Willis, Melody Schwenk, Kaitlyn Weeks, and Ruthie Ferster. 2022. Attitudes toward signing avatars vary depending on hearing status, age of signed language acquisition, and avatar type. *Frontiers in psychology* 13 (2022), 730917.
- [121] David Quinto-Pozos. 2010. Rates of fingerspelling in american sign language. In *Poster presented at 10th Theoretical Issues in Sign Language Research conference*,

- West Lafayette, Indiana, Vol. 30.
- [122] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [123] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.
- [124] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. 2021. Sign Language Production: A Review. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Nashville, TN, USA, 3446–3456. <https://doi.org/10.1109/CVPRW53098.2021.00384>
- [125] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18. Springer, 234–241.
- [126] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.
- [127] Wendy Sandler and Diane Carolyn Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge University Press.
- [128] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Adversarial training for multi-channel sign language production. *arXiv preprint arXiv:2008.12405* (2020).
- [129] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846* (2020).
- [130] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive Transformers for End-to-End Sign Language Production. <http://arxiv.org/abs/2004.14874> arXiv:2004.14874 [cs].
- [131] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Mixed signals: Sign language production via a mixture of motion primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1919–1929.
- [132] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 5131–5141. <https://doi.org/10.1109/CVPR52688.2022.00508>
- [133] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5141–5151.
- [134] Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. *arXiv preprint arXiv:2205.12870* (2022).
- [135] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA, 223–231. <https://aclanthology.org/2006.amta-papers.25>
- [136] William C. Stokoe. 1961. Sign language structure: an outline of the visual communication systems of the American deaf. 1960. *Journal of deaf studies and deaf education* 10 1 (1961), 3–37. <https://api.semanticscholar.org/CorpusID:5948293>
- [137] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018. Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *BMVC*, Vol. 2019. 1–12.
- [138] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision* 128, 4 (April 2020), 891–908. <https://doi.org/10.1007/s11263-019-01281-2>
- [139] Valerie Sutton. 1974. SignWriting. Retrieved online at: <http://www.signwriting.org/about/what/what02.html> (1974).
- [140] Garrett Tanzer, Maximus Shengelia, Ken Harrenstien, and David Uthus. 2024. Reconsidering Sentence-Level Sign Language Translation. *arXiv preprint arXiv:2406.11049* (2024).
- [141] Noam Tractinsky, Adi S Katz, and Dror Ikar. 2000. What is beautiful is usable. *Interacting with computers* 13, 2 (2000), 127–145.
- [142] Nina Tran, Richard E Ladner, and Danielle Bragg. 2023. US Deaf Community Perspectives on Automatic Sign Language Translation. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–7.
- [143] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1921–1930.
- [144] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018).
- [145] David Uthus, Garrett Tanzer, and Manfred Georg. 2023. YouTube-ASL: A Large-Scale, Open-Domain American Sign Language-English Parallel Corpus. arXiv:2306.15162 <https://arxiv.org/abs/2306.15162>
- [146] Clayton Valli and Ceil Lucas. 2000. *Linguistics of American sign language: An introduction*. Gallaudet University Press.
- [147] Adele Vogel and Jessica L Korte. 2024. What Factors Motivate Culturally Deaf People to Want Assistive Technologies?. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.
- [148] Harry Walsh, Ben Saunders, and Richard Bowden. 2024. Sign Stitching: A Novel Approach to Sign Language Production. <http://arxiv.org/abs/2405.07663> [cs].
- [149] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2024. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9326–9336.
- [150] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-Video Synthesis. *Advances in Neural Information Processing Systems* 31 (2018).
- [151] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
- [152] WFD. 2022. World Federation of the Deaf. <https://wfdeaf.org/our-work/>
- [153] Elizabeth A Winston. 1991. Spatial referencing and cohesion in an American Sign Language text. *Sign language studies* 73, 1 (1991), 397–410.
- [154] Pan Xie, Qipeng Zhang, Peng Taiying, Hao Tang, Yao Du, and Zexian Li. 2024. G2P-DDM: Generating Sign Pose Sequence from Gloss Sequence with Discrete Diffusion Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 6234–6242.
- [155] Shangqing Xu and Chao Zhang. 2024. Misconfidence-based demonstration selection for llm in-context learning. *arXiv preprint arXiv:2401.06301* (2024).
- [156] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion models: A comprehensive survey of methods and applications. *Comput. Surveys* 56, 4 (2023), 1–39.
- [157] Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. 2023. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2551–2562.
- [158] Morteza Zahedi, Daniel Keysers, Thomas Deselaers, and Hermann Ney. 2005. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Pattern Recognition: 27th DAGM Symposium, Vienna, Austria, August 31–September 2, 2005. Proceedings 27*. Springer, 401–408.
- [159] Albina Zakharenko. 2023. Semantic Differential Scale: Definition, Questions, Examples. <https://aidaform.com/blog/semantic-differential-scale-definition-examples.html>
- [160] Han Zhang, Vedant Das Swain, Leijie Wang, Nan Gao, Yilun Sheng, Xuhai Xu, Flora D Salim, Koustuv Saha, Anind K Dey, and Jennifer Mankoff. 2024. Illuminating the Unseen: A Framework for Designing and Mitigating Context-induced Harms in Behavioral Sensing. *arXiv preprint arXiv:2404.14665* (2024).
- [161] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.
- [162] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.
- [163] Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. Neural Machine Translation Methods for Translating Text to Sign Language Glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 12523–12541. <https://doi.org/10.18653/v1/2023.acl-long.700>
- [164] Jason E Zinza. 2006. Master ASL. *Sign Media Inc* (2006).

A A Review of American Sign Language and Publicly Available Datasets

Similar to other sign languages, ASL is also a visual-based natural language, expressed by using both manual and non-manual markers [136]. A common misconception is that substituting each written English word with a corresponding ASL sign would be enough as a translation [43]. However, this approach does not produce true ASL [51], as ASL has its own grammar and lexicon, distinct from English [87, 146]. Moreover, there is no one-to-one mapping between English words and ASL signs, which makes direct substitution less appropriate [102].

A.0.1 ASL Written Representation. ASL-LEX [24] has been used as a gloss reference for annotation of ASL in several works (e.g., [14, 31, 74, 90]). However, ASL-LEX glosses often lack representation of non-manual markers, such as facial expressions and body movement, which can limit the naturalness and understandability of generated signs when used in SLG [68]. To address this, ASL linguists have developed conventions to capture non-manual markers in addition to manual behaviors [100–102]. These include behaviors such as head position and movements, eye gaze and aperture, eyebrow position and movements, and body movements.

A.0.2 ASL Datasets. Sign language datasets often pose a bottleneck for SLG research [16]. Reviewing ASL datasets reveals substantial variation in vocabulary size, recording duration, number of signers, image resolution, modalities, gloss annotation conventions, and annotation tools [23, 31, 33, 34, 74, 91, 105, 106, 134, 145, 158] (Table 3). For instance, OpenASL [134] and YouTube-ASL [145] stand out with their extensive vocabularies of approximately 33,000 and 60,000 signs, respectively, offering a broad lexical base. However, these datasets provide only videos and English captions, without their corresponding written representations.

RWTH-BOSTON-50 [158] and Purdue RVL-SLLL [91] are among the earliest publicly available ASL datasets. Despite their pioneering role, their relatively small vocabularies, lack of detailed gloss annotations, non-expert human annotators, and variable image quality limit their utility for more advanced ASL research and applications. MS-ASL [74] and ASL Citizen [31] provide word-level isolated ASL signs from a wide range of signers, serving as valuable resources for sign language recognition research. However, for tasks such as generating ASL signs from English sentences, word-level datasets lack crucial contextual information, such as sentence structure, non-manual markers, and signer consistency.

Datasets like NCSLGR [106], ASLLRP [107], and DSP [105], resulting from collaborations among multiple universities, as well as the How2Sign [34] dataset collected with higher resolution cameras, offer more comprehensive data. These datasets include English sentences with corresponding written representations, detailed annotation conventions (e.g., [100, 102]), and videos featuring both continuous and citation-form signs. These advancements have allowed some of these datasets, such as NCSLGR and How2Sign datasets, to be used as benchmarks for ASL processing research (e.g., [11, 95, 163]). While these datasets address some of the critical gaps in earlier resources, issues such as their relatively small sizes (e.g., [105, 106]), inconsistent annotation conventions across datasets, and limited accessibility of the DSP and How2Sign gloss

datasets make some tasks of ASL processing both promising and challenging.

B Module 1: English Text-to-ASL Gloss

B.1 Data Preprocessing

Step 1: Data Extraction. We obtained the ASLLRP dataset from the project web interface⁵. The dataset includes ASL sentence-level signed videos and XML files⁶ containing corresponding English translations and annotations. For the translation task in Module 1, we focused on extracting manual information from the textual annotations to capture the primary meaning of the English translations. Specifically, we extracted existing English sentences from the XML files and systematically spliced English-based annotations, including vocabulary and compound symbols, fingerspelling, name signs, classifiers, locative words, and gestures, in chronological order. In total, we extracted 2,119 English sentences with corresponding English-based glosses. Additionally, we trimmed the signing videos based on the XML data so that each English sentence corresponds to a specific sign language video (utterance) for our subsequent tasks.

Step 2: Data Cleaning. Following a similar approach to prior work [3], we removed gloss annotations that did not alter the overall meaning of the sentences when omitted, such as repetition (annotated as a single or multiple “+” signs), number of signing hands (annotated as “(1h)” and “(2h)”), and signs indicator that both hands move in an alternating manner (annotated as “alt.”). To reduce translation errors, we standardized all fingerspelling-related glosses from fs-XXX to fs-X-X-X (e.g., from “fs-JOHN” to “fs-J-O-H-N”) and unified annotations for spatial locations (e.g., “i:GIVE:j” and “i:GIVE:k” were standardized to “i:GIVE:j”). While classifiers play a crucial role in ASL, we excluded them from this work because they typically appear only once or very few times in the datasets, so there was insufficient data for effective model prompting. After data cleaning, we retained 1,843 English sentences with corresponding English-based glosses for the remaining experiments.

Step 3: Text-to-Gloss Dictionary Generation. To improve consistency in sign representations across different sentences and datasets, we constructed a text-to-gloss dictionary using the ASLLRP Sign Bank⁷, which contains isolated signs along with their corresponding English-based glosses and translations. We then systematically unified the glosses based on step 2 to ensure consistency between the dictionary and the gloss annotations for the sentences. During the dictionary generation, we observed that some words may have variants of glosses depending on the context (e.g., “ask, inquire, query, question” can be annotated as “ASK”, “ASK:i”, or “i:ASK:j”, depending on whether the previous and following words are signed in a neutral location). Therefore, our dictionary employs a one-to-multiple mapping, accommodating the variability in gloss annotations. In total, the dictionary contains 3,915 text-to-gloss pairs. Notably, we identified 43 words that do not have corresponding

⁵DAI 2: <https://dai.cs.rutgers.edu/dai/s/cart>, login required.

⁶These XML files are generated from the SignStream annotation tool. More details about these files can be found here: <https://dai.cs.rutgers.edu/downloads/XML-Export-format.pdf>.

⁷<https://dai.cs.rutgers.edu/dai/s/signbank>

glosses (*i.e.*, out-of-vocab words). For these words, which lack corresponding videos, fingerspelling is used as an alternative.

Step 4: Ground True Correction. During the process of extracting ground truth from XML files to determine whether a sentence is a yes/no question, wh- question, conditional statement, and/or contains negation, we discovered that the ground truth labels were based on the signing rather than the English text, leading to some misalignments between the English text and the linguistic labels. For example, "I guarantee that the parents will be mad if the children dye their hair orange" was originally labeled as a negation statement, because the signing of it contains negation, although the English sentence does not. To address these issues, four of our researchers iteratively re-labeled and discussed the test set sentence categories, refining the labels to better reflect the text content. These revised labels were then used as the ground truth, allowing us to calculate precision and recall for each sentence type predictions and to identify patterns in the model's errors.

B.2 ASL Grammar Guidelines for LLM Prompt

American Sign Language (ASL) commonly uses a type of sentence structure called topicalization.

Topicalization is when the topic of a sentence is placed at the beginning of the sentence. For instance, in English, the topicalized form of the sentence, "I see my friend" would be "My friend, I see them". This is often referred to in ASL as topic/comment structure. Any description of the topic, such as including adjectives, would also come before the comment. The sentence "I see a big orange cat" would be signed as follows: CAT ORANGE BIG IX-1p SEE.

As a very visual language, ASL often requires signers to visualize a sentence and arrange their signs accordingly. Sentences that involve cause-and-effect statements, real-time sequencing, or general-to-specific details follow a specific pattern. Cause-and-effect sentences in ASL tend to place the cause before the effect in the sentence. For example, in the statement "I feel calm when I go to the park", the cause of "go to the park" would be expressed before the effect of "I feel calm". The sentence would be signed as: PARK GO-TO FEEL CALM ME.

Some sentences involve real-time sequencing, where events must be arranged in chronological order according to how they happened in real time.

For instance, the sentence "I'm worried because my brother didn't call me after he left" would be rearranged as: POSS-1p BROTHER LEAVE CALL-BY-PHONE-1p NOT CONCERN IX-1p.

In sentences where a signer is setting a scene, the signer should move from general to specific

details. For example, in the statement "I am excited after moving to my new house in Virginia", the signer would begin with the biggest detail ("Virginia") and work their way down to the smallest detail ("I"). The sentence would be signed as: VIRGINIA HOUSE NEW MOVE FINISH EXCITED IX-1p.

Verbs are not conjugated based on tense in ASL, so every verb is in its base form. This means that "ate", "eats", "eating", and "eaten" are all expressed by the sign EAT. The tense is established separately by including a time indicator in the sentence. Time signs are usually placed at the beginning of the sentence, before the topic, which tells the watcher when the rest of the sentence takes place. Signers can also express tense using a sign that relates the progress of the activity, like in the image above, which uses the FINISH sign to indicate that the action is in the past and translates to "I saw".

Basic sentence structure in ASL follows the pattern of Time + Topic + Comment. The word order can change depending on the needs of the signer, but this is the most common format.

- Time = Any necessary time indicators (establishes tense)
- Topic = The main focus of the sentence (a noun)
- Comment = What is being said about the topic (includes the verb)

For example, in English, one might say, "I went to the library yesterday." In ASL, the sentence might be structured like this:

- Time = YESTERDAY
- Topic = LIBRARY
- Comment = IX-1p GO-TO

As is the case with English sentence structure, sign choice and order often vary based on context. The example above is shown in Object-Subject-Verb (OSV) order, in which the object (the library) is the topic. However, the sentence can also be arranged in Subject-Verb-Object (SVO) order, in which "I" is the topic and "GO-TO LIBRARY" becomes the comment: YESTERDAY IX-1p GO-TO LIBRARY

Both sentences are grammatically correct, and different factors can influence which structure the signer chooses, such as how familiar the watcher is with the library, and therefore what level of emphasis is needed.

When a question is asked in ASL, the WHO, WHAT, WHEN, WHERE, WHY, WHICH, or HOW sign is located at the end of the sentence, or if emphasis is needed, both the beginning and the end. This word order reflects topic/comment structure. For example, in English, one might ask, "What is your name?" In ASL, the sentence would be structured in this way: YOUR NAME WHAT

Additionally, while English often employs different forms of the verb "to be" in sentences, this verb is not used in ASL and should not be included in signed conversations.

When using negating signs in a sentence, such as NOT or NONE, the negative sign typically follows the word it is negating. For example, "I don't have any pets" would be signed as: PET HAVE NOT.

B.3 Experiments on English Text-to-ASL Gloss

B.3.1 Model Selection. We experimented with various versions of GPT and tested multiple configurations to identify the optimal model. As shown in Table 5, GPT-4o-2024-05-13 (our adopted model) outperformed other GPT-4 variants under identical settings. Additionally, we fine-tuned two versions of GPT models capable of fine-tuning, but their performance was lower than that of few-shot prompting with the adopted model. However, fine-tuning GPT-4 models with larger datasets could hold promise, and exploring this option when the feature becomes more widely available may yield further improvements.

B.3.2 Prompting Examples. For Module 1, we varied the prompts for the "SYSTEM" in different setups for the English Text-to-ASL Gloss task (depicted on the left side of Figure 3), while maintaining consistency in the "ASSISTANT" and "USER" prompts. No additional prompt engineering was performed for generating linguistic information (task on the right side of Figure 3). A summary of these setups is provided in Table 6.

B.4 Additional Experiments on English-to-Gloss Translation

To enhance our translation capabilities, we implemented Retrieval Augmented Generation (RAG) [82] with anonymized embeddings. First, as a pre-process, we anonymized all train sentences by converting name references into pronouns. Next, we embedded the anonymized sentences using an OpenAI embeddings model. Finally, at inference, for each test sentence, we embedded it as well and look for the N most similar examples to this sentence based on the cosine similarity between the embedding of the test example, and the embeddings of the anonymized train examples. This way, the model is presented with the most accurate and relevant examples. As Table 7 shows, when using RAG the results are better than using all of the train examples. Moreover, using fewer examples and anonymized embeddings yields better results in most cases. The reason for using anonymization, is that names are given high

weight in the embedding, which leads to less relevant examples in some cases. For examples, the 3 most similar sentence for the sentence "Which college did Mary go to?" before anonymization, are: "Which college does Mary go to?", "What did Mary's name used to be?", "Mary used to live in Boston.". While after anonymization they are: "Which college does Mary go to?", "Which high school did you go to?", "Where did you go to high school?", which are more relevant and similar examples.

B.5 Summary of Existing Results

Unlike German datasets such as RWTH-PHOENIX-Weather 2014T [22] and the public DGS corpus [50], which are widely used and frequently reported in the literature [26, 83, 130, 157], there is comparatively less work utilizing ASL datasets. We summarize the existing translation results for ASL in Table 8.

C Survey

C.1 Section 1 (Visual and Motion Quality)

- How easy is it to understand this video? (0 = Very Hard, 1 = Hard, 2 = Neutral, 3 = Easy, 4 = Very Easy)
- If it is not easy to understand, what could be improved? (Open-ended question)
- Rate the visual quality of the signing in this video. For example, consider facial quality, blurriness. (0 = Very Poor, 1 = Poor, 2 = Neutral, 3 = Good, 4 = Excellent)
- If the visual quality is not good or excellent, what could be improved? (Open-ended question)
- Rate how natural the motion is. For example, do the movements and transitions look realistic, and are the signs produced in a typical way? (0 = Very Poor, 1 = Poor, 2 = Neutral, 3 = Good, 4 = Excellent)
- If motion quality is not good or excellent, what could be improved? (Open-ended question)

C.2 Section 2 (Translation Quality)

- Translate the ASL in this video into English. (Open-ended question)
- The intended English sentence was: "Do you have to work all night? (example)" How similar is the meaning of the video compared to the English text? (Completely Different, Not Similar, Acceptable, Similar, The Same)
- What is the quality in the ASL translation? Take into account ASL grammar and signing style but not the visual fidelity. (0 = Very Poor, 1 = Poor, 2 = Acceptable, 3 = Good, 4 = Excellent)
- How accurately does the facial expression match the English text? (0 = Very Poor, 1 = Poor, 2 = Acceptable, 3 = Good, 4 = Excellent)
- Did any of the following make the video harder to understand? (Multi-choice)
 - Grammars/sentence structure
 - Wrong signs
 - missing information
 - Wrong facial expressions
 - Lack of image clarity
 - Poor motion quality
 - Other (write below)

- There were no issues
- If you choose other, what else made it hard to understand? (Open-ended question)
- If this video does not convey the English well, how would you interpret the English sentence into ASL? Write out glosses or describe how you would sign it in ASL. (Open-ended question)

- What is your English reading and writing proficiency?(Very Poor, Poor, Acceptable, Good, Excellent)
- How often do you communicate with ASL? (Never, Monthly, Weekly, Daily)
- How often do you use spoken English? (specifically, you voicing to others) (Never, Monthly, Weekly, Daily)

C.3 Follow-up Questions and Demographics

- The following questions ask about your interest in AI Signing technology. First, imagining a version of this technology that is “nearly perfect,” meaning the videos are understandable, natural, and accurate. Answer the following with this perfect technology in mind.
 - Can you image using this technology to supplement existing live interpreters, for example they were not available or for use cases where interpreters might not be possible. (Never, Rarely, Maybe, Sometimes, Often)
 - Where might you be interested in seeing AI Signing technology? What specific applications or use cases? (Open-ended question)
 - Why are you interested in these use cases? (Open-ended question)
- All of the videos in this study are meant to look like a live ASL signer. There are alternatives, for example if the human was stylized or had a cartoon-like look. What is your interest in these styles? Assume that both versions would be capable of all signing motions needed for ASL. (All open-ended questions)
 - For what applications or purposes, if any, would you prefer video with the “live ASL signer” look?
 - Why do you think this?
 - For what applications or purposes, if any, would you prefer video that looked like a cartoon or 3D avatar?
 - Why do you think this?
- Demographics
 - What is your gender?
 - * Woman
 - * Man
 - * Non-binary
 - * Prefer not to disclose
 - * Prefer to self-describe: _____
 - What is your age range?
 - * 20 to 29
 - * 30 to 39
 - * 40 to 49
 - * 50 to 59
 - * 60 to 69
 - At what age did you learn ASL?
 - * Under age 10
 - * 11 to 20
 - * 21 to 30
 - * 31 to 40
 - * 41 to 50
 - What is your ASL understanding and production proficiency? (Very Poor, Poor, Acceptable, Good, Excellent)

Table 3: Existing ASL datasets. SL stands for sign language. “-” represents relevant information was not provided. “Unknown” represents relevant information was not found.

Dataset	Vocab.	Hours	Signers	Resolutions (pixels)	Modalities	Gloss Labeling Standard	Annotation Tools
RWTH-BOSTON-50 [158]	50	>9	3	195 × 165	Video, word	-	-
Purdue RVL-SLLL [91]	104	14	14	640 × 480	Video, Gloss	Glosses include manual English-based labels, and non-manual behaviors such as handshapes and motions for two hands.	Human Annotator
RWTH-BOSTON-400 [33]	483	-	5	648 × 484	Video, Gloss, Utterance	Glosses include manual English-based labels and non-manual behaviors, both anatomical (e.g., raised eyebrows) and functional (e.g., wh-questions). Glosses do not include handshape annotations.	SignStream@2 [100]
MS-ASL [74]	1K	24	222	224 × 224	Video, Pose, Word	Glosses were generated by referencing ASL Tutorial books [24, 164].	Human Annotator
DSP [105]	>1.7K	-	15	-	Video, Gloss, Utterance, Word	Glosses include manual English-based gloss labels, sign type, start and end handshapes (both hands), grammatical markers (e.g., questions, negation, topic/focus, conditional, relative clauses), and anatomical behaviors (e.g., head nods/shakes, eye aperture, gaze).	SignStream@3 [103]
NCSLGR [106]	1.8K	5.3	4	-	Video, Gloss, Utterance	Glosses include manual English-based labels and non-manual behaviors, both anatomical (e.g., raised eyebrows) and functional (e.g., wh-questions). Glosses do not include handshape annotations.	SignStream@2 [100]
ASLLRP [105]	>2.7K	3.6	4	-	Video, Gloss, Utterance, Word	Glosses include manual English-based gloss labels, sign type, start and end handshapes (both hands), grammatical markers (e.g., questions, negation, topic/focus, conditional, relative clauses), and anatomical behaviors (e.g., head nods/shakes, eye aperture, gaze).	SignStream@3 [103]
ASL Citizen [31]	>2.7K	30.5	52	-	Video, Gloss, Pose, Word	Glosses include manual English-based labels by referencing a lexical database of ASL (i.e., ASL-LEX [24]).	Unknown
Signing Savvy [23]	>13K	-	-	-	Video, Gloss, Utterance, Word	Glosses include manual English-based labels.	Unknown
How2Sign [34]	16K	80	11	1280 × 720	Video, Pose, Gloss, Utterance, Speech	Glosses include English-based labels, but do not include information such as hand-shape, hand movement/orientation, and facial expressions, such as raised eyebrows in yes/no questions.	ELAN [28]
OpenASL [134]	33K	288	220	-	Video, Utterance	-	-

D Gloss Annotation Conventions

Category	Gloss	Example	Explanation
English-based glosses	-	OH-I-SEE THANK-YOU	Used to separate words if the English translation of a single sign requires more than one.
	/	BOLD/TOUGH THANK-YOU	
			Used when one sign has two different English equivalents.
Fingerspelling	fs-	fs-J-O-H-N	Fingerspelled word.
	#	#EARLY	Fingerspelled loan sign.
Name Signs	ns-	ns-PARIS	Used for names of places (e.g., Paris).
Compounds	+	MOTHER+FATHER	A type of sign formation where two or more signs are joined to create a new sign with a distinct meaning (e.g., “parent”).
Phonological issues	QMwg	FRIEND FINISH DRIVE QMwg	Question marking sign (with wiggling)
Subject and object verb agreement	i:GLOSS:j	i:GIVE:j 1p:GIVE:2p	“i” and “j” designate unique spatial locations associated with the subject and object referents. “(I) give (you)...”
	Noun	fs-J-O-H-N i:GIVE:j	John is signed in a neutral location.
	Noun:i	fs-J-O-H-N:i i:GIVE:j	John is signed in the location associated with the referent (the same location with which the verb displays manual subject-verb agreement).
Agreement marking on adjectives, nouns, pronouns, determiners, possessives, and emphatic reflexives	Pronoun IX-[person]:i	IX-1p	1st person pronoun
		POSS-1p	1st person possessive marker
		SELF-1p	1st person emphatic reflexive marker (as in “I did it myself”)
	Determiner IX-3p:i	IX-2p	Pronoun referring to addressee.
		POSS-2p	Possessive marker referring to addressee.
		SELF-2p	Emphatic reflexive marker referring to addressee.

	Possessive POSS-[person]:i	IX-3p:i	Pronoun or determiner referring to singular third person referent associated with location "i".
	Emphatic reflexive SELF-[person]:i	POSS-3p:i	Possessive marker referring to singular third person referent associated with location "i".
		SELF:i	Emphatic reflexive marker referring to a singular third person referent associated with location "i".
		-	THUMB-IX-3p:i
Adverbials of location and direction	Adverbial IX-loc:[location] IX-dir:[direction]	IX-loc:i	Adverbial produced with index finger pointing to location "i".
		IX-loc"under table"	Adverbial with location described.
		IX-dir"around the corner to the right"	
		IX-loc"far"	
		THUMB-IX-loc"behind"	
Singular vs. plural	IX-[person]-[num]:i/j	IX-3p-pl-2:x/y	Third person pronoun referring to the 2 (or 3) referents: x, y (or z).
		IX-3p-pl-3:x/y/z	
		IX-1p-pl-2:x	First person pronoun referring to singer plus the referent associated with the location "i".
	-3p-pl-arc	IX-2p-pl-2:x	Second person pronoun referring to addressee plus the two referents associated with locations "x" and "y".
		IX-3p-pl-arc	Pronoun (or possessive or emphatic reflexive) referring to singular third person referent associated with location "i" articulated with the thumb.
		POSS-3p-pl-arc	
		SELF-3p-pl-arc	
1p:GIVE-3p-arc	"I give (it) to them." Subject agreement is 1st person. Object agreement (the end point of the sign) is plural (an arc).		
-loc-arc	IX-loc-arc	Adverbial ("there") using an arc to designate locations.	
Reduplicative aspect marking	Gloss-aspect	STUDY-continuative	Aspectual inflections are indicated following the gloss.
	Gloss-aspect(:i)	GIFT-distributive:i	"(they) each gave (one person)..."
Reciprocal inflection	GLOSS-recip	LOOK-AT-recip:i,j	The referents associated with locations "i" and "j" look at each other.

Table 5: Experimental results for different setups of English text-to-ASL gloss translation. Note: For "Fine-tuning," the model was not constrained to the word-to-gloss dictionary vocabulary, unlike in few-shot prompting.

Model	Training Method	Limited Vocab	Number of Examples	BLEU-4 ↑
GPT-2	Fine-tuning	-	1474 (80% of the entire dataset)	<0.000
GPT-3.5-turbo-0125	Few-shot prompting	No	100	0.102
	Fine-tuning	-	1474 (80% of the entire dataset)	0.161
GPT-4-turbo-2024-04-09	Few-shot prompting	No	100	0.115
			300	0.145
GPT-4-0125-preview	Few-shot prompting	No	100	0.117
			300	0.143
			Yes	0.176
GPT-4o-2024-05-13 (Our adopted model)	Few-shot prompting	No	100	0.133
			300	0.173
			Yes	0.226

Table 6: Prompts for different setups.

Limited Vocab	Grammar Rules	Prompts
No	No	You are an ASL translator. Your task is to translate an English sentence to an ASL gloss format.
	Yes	You are an ASL translator. Your task is to translate an English sentence to an ASL gloss format. First, familiarize yourself with the following ASL grammar rules: GRAMMER_RULES .
Yes	No	You are an ASL translator. Your task is to translate an English sentence into ASL gloss format. First, familiarize yourself the following vocabulary dictionary: TEXT_TO_GLOSS_DICTIONARY .
	Yes	You are an ASL translator. Your task is to translate an English sentence into ASL gloss format. First, familiarize yourself with the following ASL grammar rules: GRAMMER_RULES . Also, review the following vocabulary dictionary: TEXT_TO_GLOSS_DICTIONARY .

Table 7: Evaluation results of translating English text into glosses (Task on the left side in Module 1) using RAG. *All BLEU-4 and SacreBLEU scores are identical. ↑ indicates that higher values represent better performance, while ↓ indicates that lower values represent better performance. Best results in bold. The presented results are *mean ± std* across 10 repetitions of test set evaluation. Note: If “Anonymized Embeddings” is set to “No”, RAG was performed using embeddings of the original data, else, it was performed using embeddings of the anonymized data.

Number of Examples	Anonymized Embeddings	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4* ↑	ROUGE-L ↑	METEOR ↑	CHrF ↑	TER ↓
200	No	0.562 ± 0.003	0.433 ± 0.003	0.345 ± 0.003	0.278 ± 0.003	0.669 ± 0.001	0.56 ± 0.001	0.557 ± 0.002	0.52 ± 0.003
	Yes	0.569 ± 0.003	0.437 ± 0.004	0.347 ± 0.004	0.278 ± 0.003	0.668 ± 0.002	0.559 ± 0.006	0.559 ± 0.001	0.52 ± 0.002
100	No	0.563 ± 0.003	0.433 ± 0.003	0.345 ± 0.002	0.278 ± 0.003	0.663 ± 0.003	0.556 ± 0.003	0.557 ± 0.001	0.522 ± 0.004
	Yes	0.567 ± 0.003	0.437 ± 0.003	0.345 ± 0.002	0.279 ± 0.003	0.666 ± 0.002	0.562 ± 0.002	0.559 ± 0.002	0.523 ± 0.002
50	No	0.563 ± 0.003	0.432 ± 0.003	0.342 ± 0.004	0.275 ± 0.005	0.663 ± 0.002	0.557 ± 0.003	0.554 ± 0.003	0.525 ± 0.006
	Yes	0.569 ± 0.003	0.437 ± 0.003	0.348 ± 0.003	0.279 ± 0.003	0.667 ± 0.002	0.564 ± 0.002	0.558 ± 0.002	0.523 ± 0.002

Table 8: Existing English Text-to-ASL Gloss translation results reported in the literature.

References	Dataset	BLEU-4
Inan <i>et al.</i> [71]	Self-Collected ASL Dataset	0.002
Zhu <i>et al.</i> [163]	NCSLGR	0.124
Moryossef <i>et al.</i> [95]	NCSLGR	0.191