Automatic 3D animation generation for Sign Language

A case study of Sign Language Machine Translation

Luisa Martins*
University Federal of Paraíba
João Pessoa, Paraíba
luisa.lemos@lavid.ufpb.br

Stênio Ferreira
University Federal of Paraíba
João Pessoa, Brazil
stenio.ferreira@lavid.ufpb.br

Tiago Maritan
University Federal of Paraíba
João Pessoa, Brazil
tiagomaritan@lavid.ufpb.br

ABSTRACT

Digital accessibility remains a critical challenge for the Deaf community, especially in Brazil, when it comes to universalizing Brazilian Sign Language (Libras) content through different media. Traditionally, animation pipelines for Sign Language videos, including those used in the VLibras suite, are frequently slow, costly, and limited in expressiveness. This work presents a novel approach that automates the transformation of human-generated Libras videos into expressive 3D animations using deep learning-based motion transfer techniques to be applied in VLibras workflow and extend its capability in order to improve the demanded time for implementation and cost dramatically. The system leverages recent progresses in motion capture, pose estimation, and animation retargeting to enable efficient and real-time animation of human based 3D models. The present study discusses the system architecture, automation pipeline, and the potential of value creation for the Deaf community through a survey and evaluation process. Results indicate that the approach used supported the improvement on realism, enhancement on a more immersive experience, acceleration of the development and potentially high quality compared to traditional methods. Future developments on fine-tuning the model and the generation of anthropomorphic enhanced 3D models with a better collision detection capability can enable a scalable path toward a more inclusive media content for the Deaf population in Brazil.

KEYWORDS

Animation, accessibility, motion capture, Brazilian Sign Language, avatar rendering, automation

1 INTRODUCTION

In the context of the digital era, access to audiovisual and multimedia content has become not only a fundamental right but also a daily necessity. However, ensuring such access equitably remains a significant challenge for people with hearing impairments. In Brazil, data from the 2022 Demographic Census conducted by the Brazilian Institute of Geography and Statistics (IBGE) indicate that around 2.6 million people reported having some degree of hearing impairment, corresponding to approximately 1.3% of the national population [2022]. Meanwhile, according to the World Health Organization (WHO), in the "World Report on Hearing", in 2050, 2.5 billions people will be living with some degree of hearing loss, representing about 25.5% of the prospected global population (United Nations) [2017, 2013].

This significant portion of the population faces various barriers to accessing essential services such as healthcare, education, employment, and entertainment [10]. A central factor contributing to these challenges is that communication among these individuals occurs predominantly through visual-spatial languages — sign languages (SLs) — while oral languages (OLs), such as Portuguese, function only as a "second language." As a result, mediation is often required to enable communication with hearing individuals or full comprehension of written content.

In this context, some strategies, projects or initiatives have been trying in the past to enhance the accessibility in our different medias, like TV [9]. More precisely, oral language to sign language translation emerges as an essential resource, to provide visual descriptions of sound and text elements — such as speeches, conversations, and documents — enabling greater interaction, comprehension, and preservation of syntactic and cultural nuances within the scope of accessibility. However, the most traditional and widespread process for enabling this interaction is through human interpretation, which creates dependency among Deaf individuals on interpreters and restricts their access to information, especially digital media content, in their first language. Given this scenario, many assistive technology development initiatives have emerged, with machine OL-to-SL translators standing out among them.

One of the most prominent solutions in Brazil is the VLibras suite [26, 27], a government-sponsored platform that translates Portuguese text and digital content into Libras through animated avatars. The tool is already used on more than 120,000 websites — including the main portals of the federal government — and performs over 100,000 translations daily[1].

Despite its social impact and its important contribution to ensuring information access for Deaf people, as well as the technological advances achieved in this digital era, there is still a stage in the development of these translators that is predominantly manual and human-labor–based, requiring high specialization, significant time and resource investment, and facing a shortage of qualified professionals [4, 22, 25, 31].

Therefore, to improve the quality of the machine translation process and more effectively promote accessibility through this type of assistive technology, it is essential to minimize limitations, especially in the sign generation stage, such as: (i) rigid animations

*Both authors contributed equally to this research.

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

[1]The VLibras Suite is the result of a partnership between the Ministry of Management and Innovation in Public Services (MGISP), the Ministry of Human Rights and Citizenship (MDHC), and the Federal University of Paraíba (UFPB). More information: https://www.gov.br/governodigital/pt-br/vlibras/

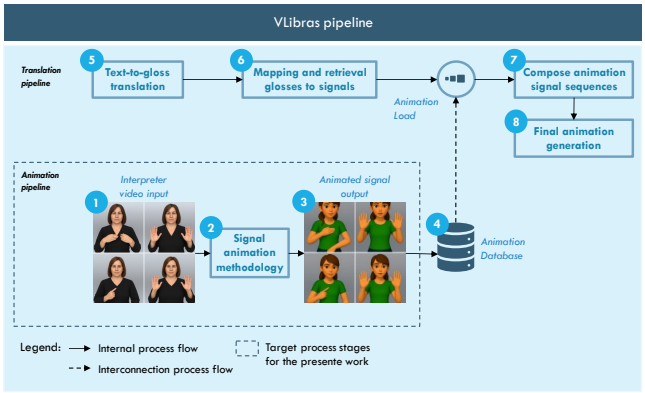

Figure 1: VLibras common execution pipeline.

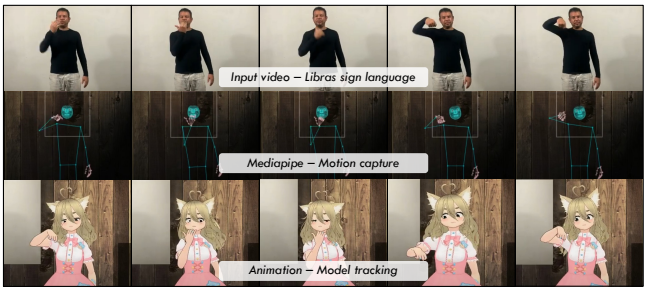

Figure 2: Animation generated by the described work, 2025

with little naturalness; (ii) the need for significant manual intervention for each animation, increasing cost and production time; and (iii) restricted facial expressiveness, which can compromise communication clarity and emotional resonance.

However, regarding advances in machine translation technologies for sign languages, one of the main bottlenecks still lies in the animation production stage, which in many systems depends on the manual creation of motion units by professional animators. This process, in addition to requiring high specialization, is slow and costly. Currently, in the conventional operation of VLibras, for example, animating a single word can require, on average, 1.6 hours of work from a professional animator. As a result, even after 15 years of operation, the system's dictionary contains around 22,000 animations, many of which exhibit low expressive variation and limited structural mobility, affecting the naturalness and realism of the signs. This limitation directly impacts scalability, restricting the quantity and diversity of available signs and contributing to the fact that SL dictionaries accumulated over the years present low expressive variation and limited naturalness in movement.

Therefore, the solution proposed in this work aims to automate the generation of sign language animations. As demonstrated in the Figure 1, a typical VLibras translation pipeline for translate text to animated sign language goes from text inputs through the design of the animated scene. Current methodologies rely on manual modeling of motion units. With the proposed process, it is expected to reduce or eliminate the average animation time per word, enabling the creation of high-fidelity animations in less time, as demonstrated in the Figure 2.

In addition to the significant time optimization, the proposed approach is expected to enhance the naturalness and expressiveness of signs, bringing them closer to the visual and kinetic characteristics observed in human performances. This improvement can contribute to increasing the quality, consistency, and scalability of animations in different contexts — such as education, public communication, healthcare services, culture, and media — making access to information in sign languages more agile, comprehensive, and faithful to the linguistic and cultural nuances of the Deaf community.

In order to evaluate the effectiveness of the proposed solution, an experimental validation was conducted with two user groups. The first group consisted of Deaf sign language consultants who were responsible to approve or reject the animations, while the second group consisted of 3D animators who had to estimate the amount of effort necessary to correct the rejected animations.

Furthermore, this proposal claims primarily for an social impact improving digital accessibility for the Deaf community. by enabling the scalable, naturalistic, and cost-effective production of sign language animations that would benefit from greater inclusion and reach. By leveraging state-of-the-art animation technologies and integrating them into existing accessibility infrastructures, such as VLibras, for example, the work seeks to transform the landscape of accessible communication in Brazil and potentially inspire similar initiatives globally.

Finally, the remainder of this article is organized in main four sections. Section 2 looks for provide a big picture of the available solutions on this field and how those solutions have impacted the this work. The next Section, 3, details the proposed solution, describing the automation framework, the computer vision model configuration, the video processing pipeline, and the specifications of the 3D models used. Section 4 presents the methodology, including the survey design, evaluation metrics, and dataset characteristics. Section 5 reports and discusses the results, highlighting the quality assessments, rejection reasons, and opportunities for process improvements. At the end, Section 6 concludes the study, summarizing the main contributions, discussing limitations, and proposing future research directions.

2 LITERATURE REVIEW

In recent years, several researchers and developers have attempted to address the issues mentioned in Section 1. Approaches include a wide range of technologies, from the creation of gesture databases [5], through the use of deep learning to classify and predict signs [3], to procedural markup animation techniques to synthesize motions algorithmically [15]. However, these methods still depend heavily on manual annotation or gloss-based translations, which fail to capture the richness of natural sign language communication. Additionally, scaling these systems across large datasets or for real-time applications remains challenging. Recent literature has explored multiple strategies to automate sign language generation, aiming to bridge communication gaps between hearing individuals and the Deaf community. One common approach is gloss-to-sign translation, where glosses — textual representations of sign language structure — are used as an intermediate step. For

example, in the article [30] was introduced a neural network-based system that translates spoken German into glosses and then into continuous sign language video using human motion data. Similarly, the author [6] proposed an end-to-end neural sign language translation model that maps spoken or written language directly into sign videos, without requiring gloss-level annotation, using the RWTH-PHOENIX-Weather 2014T dataset [6]. Another line of research focuses on avatar-based synthesis, who developed rule-based systems to animate 3D avatars performing British Sign Language (BSL) [13]. However, these models are often criticized for lacking naturalness in motion and facial expression. More recent efforts attempt to improve realism using motion capture (mocap) and pose estimation, as seen in the article [24], who use deep learning models to animate avatars with higher fidelity and expressiveness [23]. Despite progress, major limitations persist in the scalability of these systems due to data scarcity, high dependency on manual annotation, and the difficulty of modeling non-manual features (e.g., facial expressions and eye gaze), which are critical in sign languages. Data-driven sign language machine production was another approach which has been leveraged by the new advancements on computational capabilities. This approach aims to automate animation by predicting pose sequences from text or intermediate representations. Beyond transformer baselines, diffusion models now generate 3D signing motion over anatomically informed skeletons (SMPL-X), improving realism and temporal coherence while avoiding expensive manual motion libraries. Recent surveys on this topic has been consolidating these advances and highlight pose-centric pipelines as a practical route for scalable, automatic sign-language animation [21].

A notable chain of technologies developed in a huge community over the years is the ecosystem developed from HamNoSys (Hamburg Notation System) which is rule-based and with notation-driven avatars supporting to reduce animator effort by synthesizing motion directly from symbolic descriptions. Classic pipelines convert HamNoSys into SiGML and then drive an avatar such as JASigning, enabling timing control, re-use, and integration with corpus tools; recent tooling (HamNoSys2SiGML) streamlines this conversion and exposes libraries for engines like Unity, further lowering the barrier to production use. Surveys of signing-avatar animation summarize persistent challenges (non-manuals, coarticulation) and the engineering patterns that mitigate them in practice, including SiGML authoring workflows and template reuse. Projects such as DictaSign report end-to-end prototypes that translate written or speech input into avatar output with user-centred evaluation, illustrating the feasibility of scalable pipelines [11, 12, 17, 18]. A complementary system which can leverages the AZee linguistic framework to author compositional, machine-interpretable descriptions that can be rendered directly by animation systems. AZee has been used to connect structured sign language descriptions to multi-track animation, in order to synthesize productive forms beyond fixed lexicons, and to better handle spatial relocations—thereby reducing manual keyframing and improving grammatical fidelity. Recent work on synthesizing facial expressions on avatars from AZee-based inputs, improved the naturalness without hand-crafted animation passes [8].

Finally, system-level efforts combine language translation with body, hand, and facial animation to minimize manual stitching across components. A speech-to-sign avatar pipeline integrates ASR (Automatic Speech Recognition) and NMT (Neural Machine Translation), body-gesture synthesis, and facial animation into deployable applications, demonstrating that practical application in speech-to-avatar generation can be generated in a cohesive stack [19]. Over the years, these comprehensive works situate these systems within the broader translation landscape and enumerate common shortcuts, like notation reuse, pose templates and retargeting, that reduce production time consumption and rework.

3 PROPOSED SOLUTION

In the most traditional processes, to become possible the execution of the translation in sign languages, a previous effort was carried out in order to create the 3D animations as discussed in Section 2. In this context, the present work aims to create alternatives to automate and increase efficiency in an animation pipeline and supporting some links of the translation process: the sign mapping, animation synthesis, and rendering stages. The goal of this proposed solution is to reduce the lead time, minimize manual intervention, take shortcuts, streamline steps and scale up output while preserving, or improving, linguistic integrity, naturalness and expressiveness. The expected final outcome is a more agile and scalable animation pipeline that supports a wider implementation of automated animated signs in Libras in educational, governmental or corporate accessibility contexts.

3.1 Automation framework for sign language animation pipeline

The methodological backbone of this research integrates computer vision, real-time animation, and automated batch processing to automate and improve performance of an animation SL pipeline including scalable capability and expressivity. The generation of a 3D animated SL translation was carried out in three main steps (i) setup the computer vision tool - XR Animator tool; (ii) automatize video processing pipeline; and, (iii) select 3D humanoid avatars models.

3.2 Computer vision model

Table 1: Comparison of motion capture tools regarding Blender integration, cost, open-source availability, and multiple camera support

Available Tool	Blender Integration	Subscription Price	Open Source	Multiple Cameras
FreeMocap	No	Free	Yes	Yes
Remocapp	Yes	$20.00/month	No	Yes
BlendArMocap	Yes	Free	Yes	No
DollarsMocap	Yes	$99.00	No	Yes
XR Animator	Yes	Free	Yes	Yes

The computer vision model used in the current work is based on the XR Animator tool. As shown in Table 1, XR Animator stands out among motion capture solutions for combining native Blender integration, open-source licensing, and multi-camera support while remaining free of cost. These features make it particularly suitable for

scalable research pipelines without licensing constraints. Integrating MediaPipe Vision, a lightweight yet robust machine learning model that extracts 3D body landmarks per frame (BlazePose topology) with real-time inference capabilities, XR Animator provides accurate full-body pose estimation from sign videos. Originally designed to address latency and computational constraints, it enables video processing on low-capacity hardware while preserving temporal fidelity. Furthermore, leveraging the MediaPipe library, XR Animator incorporates a moving average smoothing filter (typically spanning 5–10 frames) in the extracted keypoints to reduce jitter and quantization noise [1, 2]. This temporal filtering echoes approaches used in biomechanical motion capture systems to ensure naturalistic movement without compromising linguistic articulations [16]. However, during the process, after smoothing, skeletal trajectories should be redirected onto VRM-format avatars, which is not included in MediaPipe functionalities, and extended by XR Animator.

The re-targeting pipeline implemented in XR Animator performs the alignment of source and target skeletons through a set of configurable parameters designed to ensure anatomical consistency and motion fidelity. This process involves:

- **Configurable bone scale factors:** Adjustments within a range of approximately 0.9–1.1 to match the proportions of the signer.
- **Rest position offsets:** Modifications to default joint positions, such as 10° shoulder tilt corrections, to align with natural posture.
- **Rotation constraints:** Limitation of joint rotation (e.g., ±45° per axis) to prevent anatomical distortion during motion transfer.

These re-targeting parameters also enable an enhanced biomechanical plausibility while capturing gestural nuance of Libras. Finally, resulting animations are exported in FBX or glTF formats for importing into Blender, allowing further refinement via inverse kinematics (IK) rigs and facial expression layering.

Unlike manual animation tools, this pipeline offers a reproducible and scalable approach, matching parameter-control methodologies making XR Animator a suitable choice to attend the demands of sign language gestures representation, which rely on both static posture and dynamic movement patterns.

In this project, XR Animator processes signer videos to extract skeletal poses frame-by-frame. The parameters used to fine-tune the extraction onto the model were:

- Motion capture framework: MediaPipe Vision
- AI model quality: Best
- Z-depth scale: Max

3.3 Automated video processing pipeline

In order to create scalability and high potential for maintainability, a fully automated, Python-based procedure was developed to batch-process large Libras video datasets (Figure 3). The automation cycle orchestrates sequential stages with feedback and control loops, ensuring robust processing without manual supervision. Once XR Animator has no available command interface, Selenium-driven interface automation was chosen as good automation option. Using the Selenium Python library (WebDriver), the system programmatically controls the XR Animator GUI—positioning the mouse, clicking buttons, handling menus, launching export commands, and validating completion status in a headless or visible browser context:

- **Batch ingestion:** The pipeline continuously monitors designated input folders, ingesting new video files as they appear.
- **Avatar targeting:** normalized inputs are fed into XR Animator, which maps control points and movements over time to a pre-configured virtual avatar skeleton.
- **Video rendering:** once the motion capture is carried out, then rendering is triggered, producing final synchronized output videos ready for distribution.

A feedback loop monitors each step's success or failure: failures trigger retries or alerts; successes are logged for auditing and performance metrics. This automated cycle significantly reduces human overhead—where previously manual intervention was required for GUI operations in XR Animator, now everything proceeds in batches unattended. It enables high-throughput processing of hundreds or thousands of clips while preserving consistency and traceability.

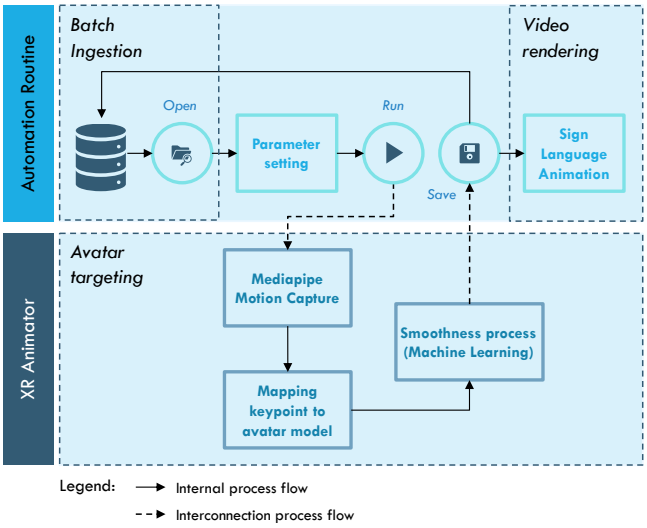

Figure 3: Automation of animation generation process.

3.4 3D avatar model specification

The 3D avatar used in this work is a humanoid VRM-format model. These models are used in applications of virtual reality and streaming, and generally are chosen for its expressiveness capabilities (with its facial control shape) and real-time performance compatibility. The VRM format is a binary file, glTF-based, platform-independent standard that encapsulates not only mesh geometry but also a normalized humanoid skeleton, blendshape groups (morph targets), secondary animation data (e.g. spring-bones), material definitions, eye-gaze metadata and licensing/authoring metadata packaged into a single file [33]. It supports blendshape animation, hand rigging,

and lip sync features, allowing a natural presentation of Libras signs including facial expressions and handshapes.

A typical VRM avatar comprises a hierarchical bone structure matching a T-pose skeleton—spine, chest, neck, head, upper/lower arms, hands, fingers, legs, and other additional parts with joint orientation standardized to permit animation retargeting conformant. Facial expressiveness is carried out via blendshape groups that encode pre-defined emotions and phonemic mouth shapes (e.g. ARKit-style visemes for "a-i-u-e-o", blink, joy, anger, surprise) VRM. These blendshapes rely on control points defined at the vertex level (morph targets) to shape facial regions with precise weights. XR Animator is capable of execute the orchestration, normalized keypoints from MediaPipe are matched to this skeleton: body and hand joints drive joint rotations, while facial landmarks drive blendshape activations for lip sync and expressions.

During empirical testing, we leveraged large VRM model repositories like [14, 20, 29], which are generally not specific for sign language applications. For this reason, some limitations emerged when using those avatars with non-standard anthropometry proportions or low-resolution mesh structure. Models with exaggerated limb lengths or stylized proportions caused misaligned retargeting: finger positions, palm orientations, or facial expressions became distorted or inconsistent with Libras sign shapes in some cases. The low-precision collision meshes (capsule colliders or basic rigid bodies) also led to avatar self-intersection or mesh clipping during complex hand gestures, reducing visual realism. In some cases, incompatible blendshape bases - due to malformed or incorrectly-exported VRM files - resulted in abnormal mesh deformation (e.g. stretched facial features). Despite these constraints, the chosen VRM avatar provided a practical balance between fidelity and computational performance for a proof of concept model, allowing efficient real-time rendering of fluent Libras animations with synchronized facial and manual expression. However, parameter fine-tuning and using avatars with precise dimensions calibration and higher-resolution blendshape bases should minimize distortions and preserve sign clarity.

4 METHODOLOGY

To validate the proposed solution presented in Section 3, a set of experiments was conducted with deaf sign language users and professional 3D animators. The evaluation consisted of three phases. The first phase was conducted with deaf sign language users and aimed primarily to determine whether the generated sign was performed correctly and would be approved for inclusion in a sign dictionary. The second phase of the experiment was carried out with 3D animators, who were presented with the signs rejected by the deaf consultants. The animators were then asked to indicate the level of effort required to adjust the sign and the necessary corrections. The third phase consisted of analyzing the results obtained. Finally, it is important to highlight that the project was submitted and approved by the Research Ethics Committee of Federal University of Paraíba under Opinion No. 6.329.894 (CAAE 72907123.6.0000.5188). All participants were informed about the purpose of the research and signed the *"Termo de Consentimento Livre e Esclarecido"* (TCLE).

Figure 4: Example of VRM models tested and its features of blendshapes and dimensions.

4.1 Survey methodology & evaluation metrics

In this work, interpretability of the results was assessed by a survey involving 7 participants (5 male, 2 female), aged between 20 and 30 years, all with completed secondary education and currently enrolled in undergraduate or postgraduate programs. The sample comprised 3 deaf consultants who assessed the animation quality and 4 professional animators who rated the correction effort. The objective consisted of running an online survey with multiple-choice, Likert-scale, and free response questions. The survey took about 30 minutes.

An evaluation survey was designed with questions regarding the expressiveness, legibility, timing, and realism of the animations. Deaf consultants were invited to assess the automatic-generated animations of SL signs corresponding to OL words ($n = 100$) in the context of education and health. In addition to quantify it was created a 1–5 scale, open-ended qualitative feedback, and finally participants were also asked whether each animation would be approved or rejected in a formal accessibility review. Subsequently, animations marked as "rejected" were evaluated by the team of professional animators, who rated the amount of effort required to fix each animation on a scale from 1 (minimal adjustments) to 5 (complete rework). The questionnaire was composed of the following profile of items:

(1) **Do you approve this animation?:** *Yes/No option — captures binary user endorsement of the animation quality.*
(2) **If you disapproved the animation, what were the reasons?:** *Closed-ended question — allows participants to choose among categoric answers. (i) Error in hand movement; (ii) Inadequate facial expression; (iii) Occlusion hindering visualization; (iv) Lack of fluidity; (v)Mesh penetration (avatar geometry intersecting itself)*
(3) **Overall quality of the animation:** *Likert-type item (5-point scale: Very Poor to Excellent) — measures holistic perception of production quality and clarity.*
(4) **Number of hands used in the sign:** *Ordinal scale (e.g., One hand, Two hands) — checks whether the hand configuration matches the expected gloss morphology.*
(5) **Is there occlusion? (i.e., are parts of the body or hands hidden?):** *Yes/No option — evaluates visibility issues where hand shapes or body parts may be obscured.*
(6) **Does the animation present appropriate facial expression?:** *Yes/No option — assesses synchronization of non-manual markers (facial cues) critical to Libras comprehension.*
(7) **Additional comments:** *Open-ended field — provides space for qualitative feedback and suggestions beyond structured items.*

In the further assessment with specialized SL designers, it was addressed a survey to test the effort of correct the animations:

(1) **What level of effort was required to correct the animation?:** *Likert-type item (5-point scale: Very Low to Very High) — measuring perceived effort or difficulty in usability and evaluation studies.*

4.2 Dataset and resources

The dataset used comprises video clips featuring professional SL interpreters recorded in medium shot under varied lighting and environmental conditions and is called $VLibras-Sign-v1$. These clips were provided and originally produced by the VLibras community, with explicit consent for usage, within the scope of institutional accessibility and inclusion efforts. Each video corresponds to a single sign from a curated vocabulary related primarily to health and education. This dataset offered a solid foundation for evaluating motion capture consistency and avatar expressiveness.

In the context of sign language video production, the referred medium shot, or American shot, is a cinematographic framing technique in which the subject is captured from approximately the waist up, facing the camera in a frontal or slightly angled orientation. This framing has become particularly suitable for sign language recording due to its ability to simultaneously capture the full articulation of the upper limbs and the detailed facial expressions. This visual composition is critical because SL relies heavily on non-manual markers, including eyebrow movement, mouth shapes, head tilts, and eye gaze, which convey essential grammatical and emotional cues. At the same time, the signer's handshapes, movements, and positions relative to the body must remain clearly visible to ensure lexical and syntactic intelligibility to the capture. By using the $VLibras - Sign - v1$ dataset, it has been trying to maximize the semantic content carried in both manual (hands, arms) and non-manual (face, shoulders, upper torso) components of the sign. Moreover, this shot allows for a neutral background and consistent lighting, which enhances pose estimation accuracy in automated animation pipelines as MediaPipe.

Table 2: Dataset description

Field	Description
Title-Version	VLibras-Sign-v1
Description	24660 videos from 11 signers and 406 signs (206 educational and 200 health related signs)
Creators	Lavid UFPB
Formats	MP4, 1280x720, 1920×1080, 30 fps, ~2-4 MB per video
Demographics	4 male / 7 female signers, ages 18–45, regional backgrounds (Northeast Brazil)
Collection setup	Unique medium-shot framing, frontal uniform lighting, neutral background
Licenses	Private dataset

Figure 5: Dataset features: setup and interpreters of SL.

5 RESULTS AND DISCUSSION

5.1 Survey descriptive results

Deaf participants first reviewed a sample of animations and made accept/reject judgments, while animators later evaluated the effort required to refine those animations. From the collected judgments of two groups of participants, it was possible to infer that from all automatically generated animations evaluated by the deaf consultants, only 38% were approved, while 62% were rejected, as shown in

the Table 3. Besides that, the quality ratings for rejected videos cluster around 3 ("Regular") in the Table 4, suggesting that a significant portion was considered technically acceptable but not satisfactory enough for reach an approval standard. This intermediate quality category represents a potential opportunity for refinements, rather than wholesale reconstruction. On the other hand, although these findings were associated with the majority of animations, most corrections required minimal effort, according to professional animators Table 5. Specifically, approximately 77% of videos fell into the very low to low correction-effort categories (levels 1 and 2), indicating that even rejected animations were relatively easy to fix, and showing an optimization level.

Analyzing the reasons for the rejection, Figure 6, it was possible to understand the low performance animation problems. From a total of unapproved animations $n = 186(62\%)$, the most relevant impacts were recorded in: (i) hand movement errors ($p = 84.7\%$); (ii) occlusion that hinders visibility ($p = 27.3\%$); (iii) mesh intrusion or intersection ($p = 17.5\%$). These 3 categories of rejection reasons are present in $p = 96,7\%$ of the unapproved animations. It was also questioned about the animation quality in the case of unapproved sign-language translation production. Despite not achieving full approval, the animation quality was rated as "Good" (rating of 4) or "Very Good" (rating of 5) in 31% of the videos exhibiting hand movement errors, and in 50% of those experiencing occlusion that hinders visibility. Finally, the feasibility to improve the animation was assessed by the qualitative effort demanded by a graphical designer in reconfiguring the animation in order to have a acceptable result, as demonstrated by the Figure 7. For the recorded answers with: (i) hand movement errors ($\theta = 76,6\%$); (ii) occlusion that hinders visibility ($\theta = 78,9\%$); (iii) mesh intrusion or intersection ($\theta = 78,9\%$); demonstrated a "Low" or "Very low" pointed effort to correct the animations, which the authors considers as a good result for the proof of concept proposed in this article.

Table 3: Automatic generated animations evaluation by deaf consultants

Evaluation	Percentage of videos	n samples
Approved	38%	114
Rejected	62%	186
Total	**100%**	**300**

Table 4: Quality evaluation of rejected automatic generated animations by deaf consultants

Quality	Percentage of videos	n samples
1 - Too bad	11%	20
2 - Bad	24%	45
3 - Regular	35%	65
4 - Good	17%	32
5 - Very good	13%	24
Total	**100%**	**186**

Table 5: Amount of effort necessary to fix the animations evaluated by animators

Effort	Percentage of videos	n samples
1 - Very low	39%	73
2 - Low	38%	70
3 - Medium	21%	39
4 - High	2%	4
5 - Too high	0%	0
Total	**100%**	**186**

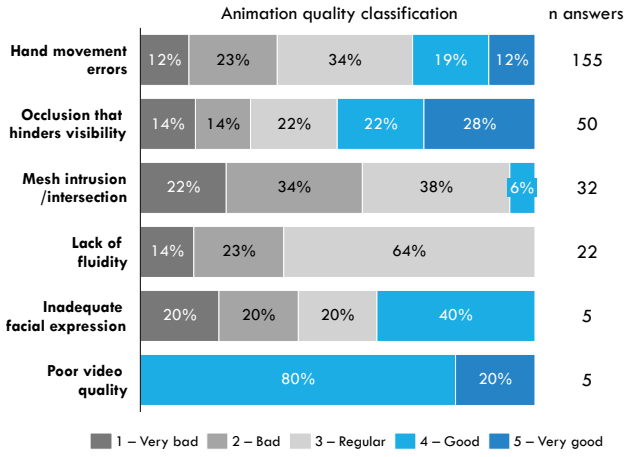

Figure 6: Video quality in comparison with main reasons for the rejection of animations.

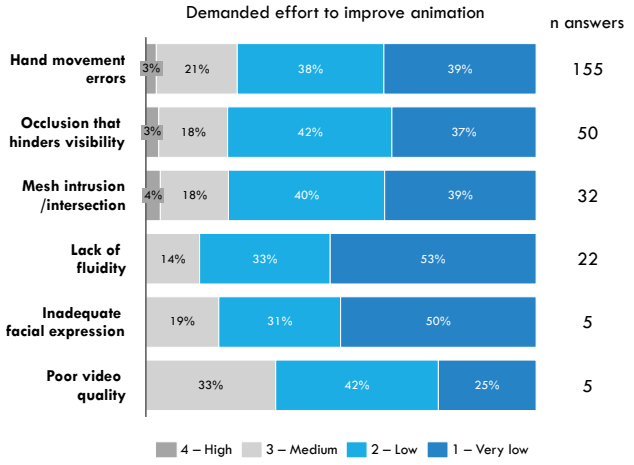

Figure 7: Effort demanded to improve animation according to main reasons for the rejection of animations.

5.2 Survey qualitative results

As the main reasons for rejection can present a wide range of possible causes, they demand to be deeper analyzed in order to

create future mechanism which mitigate the most frequent failures in the sign-language process animation generation, as demonstrated in the Figure 8. An individual analysis over the hundred samples presented in the survey, it was possible to identify root causes or limitations classified into two steps over the generation process: (i) 3D avatar mapping and retargeting; (ii) MediaPipe motion capture.

As demonstrated in the Figure 8(a), the right hand shows a penetration in the hair and head parts of the 3D avatar, hiding part of the sign and preventing from a correct visualization. In these mesh intrusion problems, the collision and mesh penetration presented mainly in the avatar mapping and retargeting step were predominant in the samples. In the Figure 8(b), a case of hand movement misalignment between the avatar or MediaPipe and the input video is shown as a wrong hand movement and head height position, where these deviations led to anatomically inconsistent arm placement and consequently incorrect interpretation. The authors recognize that the MediaPipe process demonstrates to be more precise [28], leading up to the occurrence mostly in the avatar retargeting process. A more complex and multi-factorial situation is shown the Figure 8(c), which was observed a simultaneous occurrence of hand occlusion and arm mispositioning. The primary cause of hand occlusion was identified as a limitation of the MediaPipe framework in accurately capturing overlapping movements, particularly when one hand passes in front of the other. This constraint in pose estimation often results in partial or complete loss of keypoint tracking for the occluded hand. In parallel, the arm mispositioning was found to originate within the avatar retargeting stage. The combination of these two errors adversely affects the clarity and naturalness of sign representation.

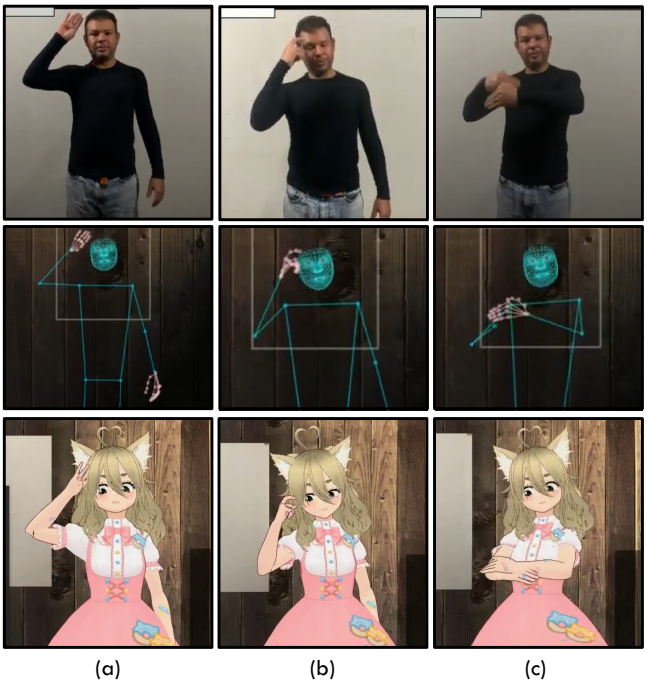

Figure 8: Most common errors in sign-language translation: (a) Mesh invasion; (b) Hand movement errors; and (c) occlusion.

5.3 Limitations

Despite the promising outcomes achieved by the proposed system, several limitations were identified throughout the execution and the survey. The motion capture framework (MediaPipe) demonstrates to struggle with hand occlusions and overlapping gestures, which may lead to partial keypoint loss and inaccurate trajectory reconstruction. During the retargeting process, the use of non-standard or low-resolution VRM avatars occasionally produced mesh penetration and anatomical distortions. For the survey, a small number of participants with a relative high variance prevent the work from further significant hypothesis testing on methodology effectiveness.

6 CONCLUSIONS AND FUTURE WORKS

This study introduces a novel method to automatically animate human-recorded Libras videos into expressive 3D avatar animations using deep learning-based motion transfer techniques. The core idea is to bypass traditional animation methods, optimize and improve efficiency in a pipeline of sign language animation. The system captures real human gestures and delivers high-fidelity animations with minimal manual intervention. Beyond technical contribution, the current work proposes an approach that addresses critical barriers in accessibility workflows by reducing production costs, accelerating content creation, and ensuring that the automatic generated animations preserve the linguistic accuracy and expressiveness essential for sign language comprehension. The potential of scalability and integration with platforms such as VLibras positions it as a viable method toward the primary objective of this work: a social impact from a broader dissemination of accessible content for the Deaf community in Brazil.

From its findings, the study revealed the clear need of improved motion capture algorithms capable of handling occlusion and more precise avatar rig calibration to ensure biomechanical plausibility in which code-based automation can mitigate the most frequent failures in the sign-language process animation generation, as demonstrated in Figure 8. These strategies in general address rendering of avatar and its control points programmatically like kinematic smoothing algorithms, motion interpolation, and collision-aware skeleton constraints, which can be implemented in code to enhance fluidity, minimize mesh intrusion, and correct hand trajectory deviations in post-production code efforts. For example, future works will need to tackle the case of hand movement misalignment. Those irregularities can be addressed by scripting or procedural corrections via code, which aim to refine standard control parameters or manipulate bone pose angles in order to enhance the performance. Other solutions like post-processing filters in code to smooth joint trajectories can significantly improve accuracy. Analogously, mesh intrusion problems can be faced with a enhanced collision detection calibration process avoiding a better underlying rig demand. Finally, for inadequate facial expression integrating facial grammar modules into the animation pipeline can enrich the quality of identification of correct facial expressions based on gloss enhancing expressiveness.

Future work will focus on enhancing finger and facial tracking precision, creating higher quality 3D models with improved anthropomorphic proportions and more accurate collision detection, fine-tuning smoothness and keypoint mapping parameters, and developing a better understanding of the root causes for animation rejection. Additional efforts include creating open datasets and benchmarks for sign language animation, expanding the approach to other sign languages and multilingual contexts, and improving integration into the VLibras workflow to broaden testing and adoption of the tool.

With this work, the authors aim to contribute to the recent growth in the adoption of new technologies in the accessibility fields for Deaf communities.

7 ACKNOWLEDGEMENTS

Acknowledgments to the *"Ministério de Gestão e Inovação em Serviços Públicos"* (MGI) of the Brazilian Federal Government, through the *"Secretaria de Gestão e Inovação"* (SGI), for funding this research.

REFERENCES

[1] 2021. Body Posture Detection & Analysis System using MediaPipe. LearnOpenCV tutorial. Describes BlazePose topology and real-time 3D landmark detection.

[2] 2025. Pose landmark detection guide. Google AI Edge documentation. Accessed online, describes MediaPipe Pose capabilities for 33 3D landmarks.

[3] T. N. Abu-Jamie. 2022. Classification of Sign-Language Using MobileNet. *International Journal on Recent and Innovation Trends in Computing and Communication* (2022).

[4] Bader Alsharif, Easa Alalwany, Ali Ibrahim, Imad Mahgoub, and Mohammad Ilyas. 2025. Real-Time American Sign Language Interpretation Using Deep Learning and Keypoint Tracking. *Sensors* 25, 7 (2025). https://doi.org/10.3390/s25072138

[5] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. RWTH-PHOENIX-Weather 2014T: Parallel corpus of sign language video, gloss and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT, USA. https://doi.org/10.1109/CVPR.2018.00812

[6] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Roberto Cavararo. 2022. *Censo 2022 – Pessoas com Deficiência e Pessoas diagnosticadas com Transtorno do Espectro Autista (TEA)*. Instituto Brasileiro de Geografa e Estatística (IBGE), Rio de Janeiro. 211 pages. https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/43463-censo-2022-brasil-tem-14-4-milhoes-de-pessoas-com-deficiencia

[8] Camille Challant and Michael Filhol. 2022. A First Corpus of AZee Discourse Expressions. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*. 1537–1546. https://aclanthology.org/2022.lrec-1.167.pdf

[9] Richelieu R. A. Costa, Derzu Omaia, Tiago M. U. Araújo, Jóison O. Pereira, Anderson S. Coutinho, Miguel P. S. Cruz, Victoria M. Pontes, Matheus M. Barbosa, Abner S. Silva, and Guido L. S. Filho. 2023. Acessibilidade na TV 3.0 Brasileira a partir de mídias de legenda, glosa e áudio descrição. In *Anais Estendidos do Workshop Futuro da TV Digital Interativa / Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*. Sociedade Brasileira de Computação, Porto Alegre, Brasil, 123–129. https://doi.org/10.5753/webmedia_estendido.2023.236168

[10] Maria Fernanda Neves Silveira de Souza, Amanda Miranda Brito Araújo, Luiza Fernandes Fonseca Sandes, Daniel Antunes Freitas, Wellington Danilo Soares, Raquel Schwenck de Mello Vianna, and Árlen Almeida Duarte de Sousa. 2017. Principais dificuldades e obstáculos enfrentados pela comunidade surda no acesso à saúde: uma revisão integrativa de literatura. *Revista CEFAC* 19, 3 (jun 2017), 395–405. https://doi.org/10.1590/1982-0216201719317116

[11] Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Lefebvre-Albaret. 2012. The Dicta-Sign Wiki: Enabling Web Communication for the Deaf. In *Computers Helping People with Special Needs (ICCHP 2012)*. Lecture Notes in Computer Science, Vol. 7383. Springer, 205–212. https://doi.org/10.1007/978-3-642-31534-3_32

[12] Ralph Elliott, John R. W. Glauert, Vicki Jennings, and Richard Kennaway. 2004. An Overview of the SiGML Notation and SiGMLSigning Software System. In *Proceedings of the LREC 2004 Workshop on the Representation and Processing of Sign Languages*. ELRA, Lisbon, Portugal, 98–104. https://www.sign-lang.uni-hamburg.de/lrec/pub/04020.pdf

[13] Ralph Elliott, John R. W. Glauert, Richard Kennaway, Ian Marshall, and Éva Sáfár. 2008. Linguistic modelling and language-processing technologies for Avatar-based sign language presentation. *Universal Access in the Information Society* 6, 4 (2008), 375–391. https://doi.org/10.1007/s10209-007-0102-z

[14] HTC Corporation. 2025. VIVERSE: Platform supporting import of VRM avatars. https://avatar.viverse.com/. Accessed July 2025.

[15] Richard Kennaway. 2015. Avatar-independent scripting for real-time gesture animation. In *arXiv preprint*. Introduz o SiGML, linguagem de marcação para gerar animações de sinais de forma procedural e independente do avatar.

[16] Jong-Wook Kim, Jin-Young Choi, Eun-Ju Ha, and Jae-Ho Choi. 2023. Human Pose Estimation Using MediaPipe Pose and Optimization Method Based on a Humanoid Model. *Applied Sciences* 13, 4 (2023), 2700. https://doi.org/10.3390/app13042700

[17] Carolina Neves, Luísa Coheur, and Hugo Nicolau. 2020. HamNoSyS2SiGML: Translating HamNoSys Into SiGML. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. ELRA, Marseille, France, 5991–5998. https://aclanthology.org/2020.lrec-1.739.pdf

[18] Aritz Núñez-Marcos, Xabier Alameda-Pineda, and Elisa Ricci. 2023. A Survey on Sign Language Machine Translation. *Expert Systems with Applications* 213 (2023), 118993. https://doi.org/10.1016/j.eswa.2022.118993

[19] Adrián Núñez-Marcos, Olatz Perez de Viñaspre, and Gorka Labaka. 2023. A survey on Sign Language machine translation. *Expert Systems with Applications* 213 (2023), 118993. https://doi.org/10.1016/j.eswa.2022.118993

[20] Pixiv Inc. 2025. VRoid Hub: repository and API for VRM avatars. https://vroid.com/. Accessed July 2025.

[21] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, Vassilis Athitsos, and Mohammad Sabokrou. 2024. A survey on recent advances in Sign Language Production. *Expert Systems with Applications* 243 (June 2024), 122846. https://doi.org/10.1016/j.eswa.2023.122846

[22] Len Roberson and Sherry Shaw. 2024. Signed Language Interpreter Education Programs in North America: A Descriptive Study. *Journal of Interpretation* 32, 1 (2024), Article2. https://digitalcommons.unf.edu/joi/vol32/iss1/2

[23] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[24] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. *International Journal of Computer Vision* (2021).

[25] Nada Shahin and Leila Ismail. 2024. From rule-based models to deep learning transformers architectures for natural language processing and sign language translation systems: survey, taxonomy and performance evaluation. *Artificial Intelligence Review* 57, 10 (2024), 271–351. https://doi.org/10.1007/s10462-024-10895-z

[26] Bruno Cassol Silva. 2023. VLIBRAS E Governo Digital: uma análise da ferramenta eletrônica e das variações linguísticas da Libras. *Revista Brasileira de IA e Direito (RBIAD)* 1, 2 (2023). https://rbiad.com.br/index.php/rbiad/article/view/5 IV Mostra de Reviews, Cases e Insights do IV SIAD.

[27] Luana Silva, Tiago Maritan U. Araújo, Maria Dayane F. Cirino Lima, Angelina S. da Silva Sales, and Yuska Paola Costa Aguiar. 2017. Avaliação de Usabilidade do Aplicativo VLibras-Móvel com Usuários Surdos. In *Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*. Sociedade Brasileira de Computação, Porto Alegre, Brasil, 123–126. https://doi.org/âĂŞ

[28] W SIMOES, L. REIS, C. ARAUJO, and J. MAIA JR. 2024. Accuracy Assessment of 2D Pose Estimation with MediaPipe for Physiotherapy Exercises. *Procedia Computer Science* 251 (2024), 446–453. https://doi.org/10.1016/j.procs.2024.11.132 15th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 14th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare EUSPN/ICTH 2024.

[29] Sketchfab Inc. 2025. Sketchfab: Platform for 3D models, VRM avatars via WebGL/WebXR. https://sketchfab.com/. Accessed July 2025.

[30] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2018. Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *Proceedings of the British Conference on Machine Vision (BMVC)*. BMVA Press, Newcastle, UK.

[31] Nina Tran, Richard E. Ladner, and Danielle Bragg. 2023. U.S. Deaf Community Perspectives on Automatic Sign Language Translation. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 76, 7 pages. https://doi.org/10.1145/3597638.3614507

[32] United Nations. 2017. *World Population Prospects: The 2017 Revision*. Technical Report. United Nations Department of Economic and Social Affairs. https://population.un.org/wpp/ Accessed: 2023-10-09.

[33] VRM Consortium, Inc. 2019. VRM: a humanoid 3D avatar file format based on glTF 2.0. https://vrm-consortium.org/en/. Platform-independent avatar standard

supported by UniVRM.
[34] WHO World Health Organization. 2013. *Millions of people in the world have hearing loss that can be treated or prevented.* WHO Document Production Services, Geneva. https://www.who.int/publications/i/item/9789241506571