Product Requirements

Target release	25 May 2025
Epic	Develop, Test and Deploy MLOps (MLOps level 2)
Document status	DRAFT, V0.2
Product Owner	Hoang Quan Dang
Solution Designer	
Tech Lead	
Stakeholders	

Objective

This project aims to enhance communication accessibility for deaf and hard-of-hearing community by providing real-time translation of continuous American Sign Language (ASL) into accurate English text with <2s latency, integrated into video conferencing platforms like Google Meet. It targets inclusivity in communication scenarios such as virtual meetings, educational settings, and public services, where deaf individuals often face barriers due to the lack of interpreters. As part of the "AI for Good" initiative, this product leverages Computer Vision and Natural Language Processing (NLP) to bridge the communication gap between signers and non-signers without requiring a particular hardware, physical gloves, or glasses, relying only on standard webcams. The project aligns with the Accessibility and Inclusion Innovation Team's broader mission to enhance digital accessibility and promote social inclusion, directly supporting the United Nations (n.d.)'s Sustainable Development Goal 10 (Reduced Inequalities). By enabling seamless communication between deaf ASL users and non-signers in virtual settings (e.g., education, workplaces, social interactions), this project will:

- Facilitate participation of deaf individuals in virtual professional, educational, and social settings.
- Drive innovation by showcasing advanced AI applications (Computer Vision, NLP, MLOps) in accessibility solutions.
- Reduce communication barriers for the deaf community, encouraging equity in digital communication.
- Directly benefit an estimated 1 million ASL users, according to (National Geographic, 2024).

Definitions

- **Sign Language**: A visual means of communication using hand gestures, facial expressions, and body movements to convey meaning, which is primarily used by deaf and hard-of-hearing individuals. Unlike spoken language, sign Language has its own grammar and syntax (e.g., facial expressions for questions, negation). American Sign Language (ASL), a distinct sign language used primarily in the United States and Canada, is one example of more than 300 sign languages in the world (National Geographic, 2024).
- **Continuous sign language translation**: Unlike isolated sign recognition (e.g., fingerspelling alphabets), continuous translation involves interpreting a sequence of signs, including transitions, grammar (e.g., raised eyebrows for questions), and context, to produce coherent sentences.
- **Gloss**: A written representation of an individual sign (e.g., "MOTHER", "WANT"), typically in uppercase. Glosses are intermediate outputs in the translation pipeline, capturing individual signs before conversion to natural language.
- **Pose Estimation**: A computer vision technique to detect and track key points on a human body (e.g., hand joints, facial landmarks) from video frames.
- **Key Points**: Specific body landmarks extracted from video, including 21 hand joints per hand, 468 facial landmarks, and 33 pose landmarks, used to interpret signs.

- **Pose-to-Gloss**: The process of mapping a sequence of key points (hand, body, or facial landmarks extracted from video frames) to corresponding glosses using a Machine Learning model (e.g., Transformer). This step captures the semantic units of ASL.
- **Gloss-to-Text**: The process of translating a sequence of glosses into natural, grammatically correct sentences (e.g., English's subject-verb-object order) using an NLP model (e.g., GPT).
- **BLEU Score**: A metric to evaluate translation quality by comparing machine-generated text to reference translations.

Assumptions

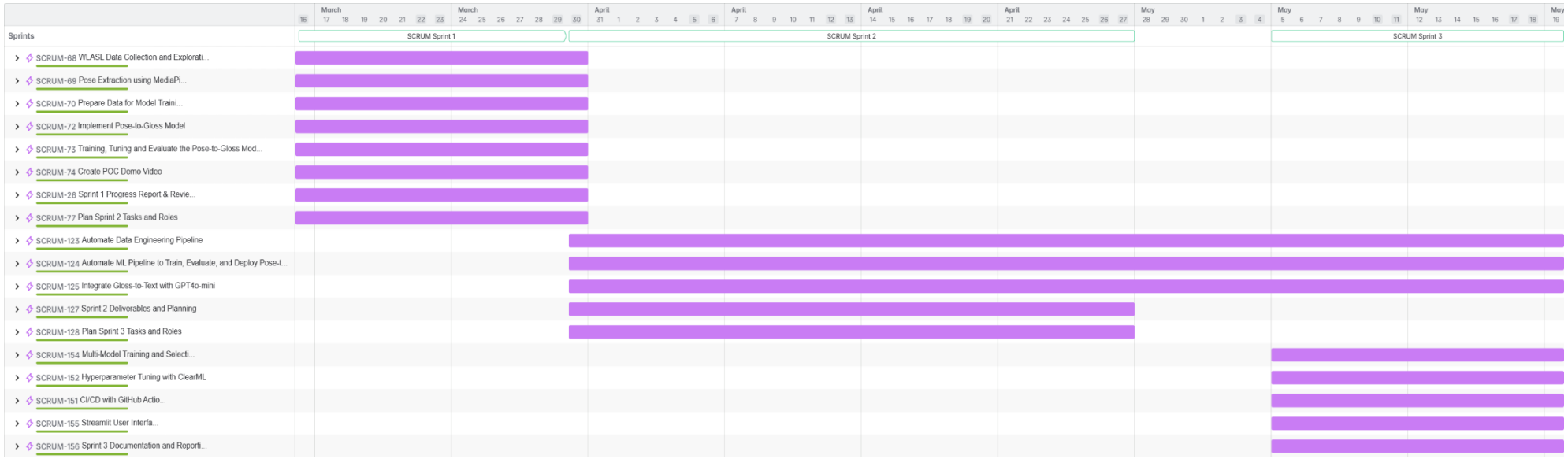
1. The AI model will accurately recognise and transcribe sign language with 80% of accuracy in real-time
2. They system will initially support ASL, with future updates expanding coverage of more sign language.
3. The AI Model will continuously improve through ML updates based on user feedback and real-world interactions.
4. Users (deaf individuals, educators, non-signers) have access to devices with a standard webcam (720p, 30 FPS) and a modern web browser (e.g., Chrome, Firefox).
5. Users are familiar with conferencing platforms like Google Meet and can engage with the project's interface.
6. Non-signing users understand English as the output language and Deaf users performing ASL have basic proficiency. This aligns with Barnett et al. (2011)'s study, which estimates that 100,000 to 1 million people use ASL as their primary language.
7. Input videos are recorded in well-lit conditions with minimal background noise or occlusions and contain a single signer performing ASL. This assumption aligns with MediaPipe's performance, which can achieve high pose estimation accuracy in controlled conditions (Dill et al., 2023).

8. The dataset we aimed to use, WLASL (Li et al., 2020), which contains 2,000+ ASL signs performed by more than 100 signers, making it a standard benchmark for ASL recognition tasks, is sufficient for training and evaluating the Pose-to-Gloss model.
9. Real-Time Performance with <2s latency is feasible with optimized models and standard hardware (e.g., 8-core CPU, 16GB RAM). Dice and Kogan (2021) show a potential where Transformer optimization (e.g., quantization) results in 2.37x speedup in BERT inference on CPUs.
10. Cloud platforms (e.g., Google Colab Pro+ with A100 GPUs) are available for training Transformer models. This is supported by typical academic access to such resources. Training a Transformer with 4 layers and 8 heads on WLASL typically may require ~10 hours on a V100.
11. The project will be adopted by accessibility-focused organizations (e.g., universities, NGOs) post-deployment. World Health Organization (2021) estimates that over 1.5 billion people worldwide have disabling hearing loss, a figure projected to reach 2.5 billion by 2050, driving demand for accessibility solutions.

Success metrics

Business Goal	Definition	Quantitative Metric
The system will achieve at least 80% of top 5 accuracy in real-time sign language recognition with <2s latency on consumer hardware.		
The user retention rate should be above 70% after 3 months of usage.		
The system should integrate with major platforms such as video conferencing tools, media platforms, and workplace apps.		

Milestones



Requirements

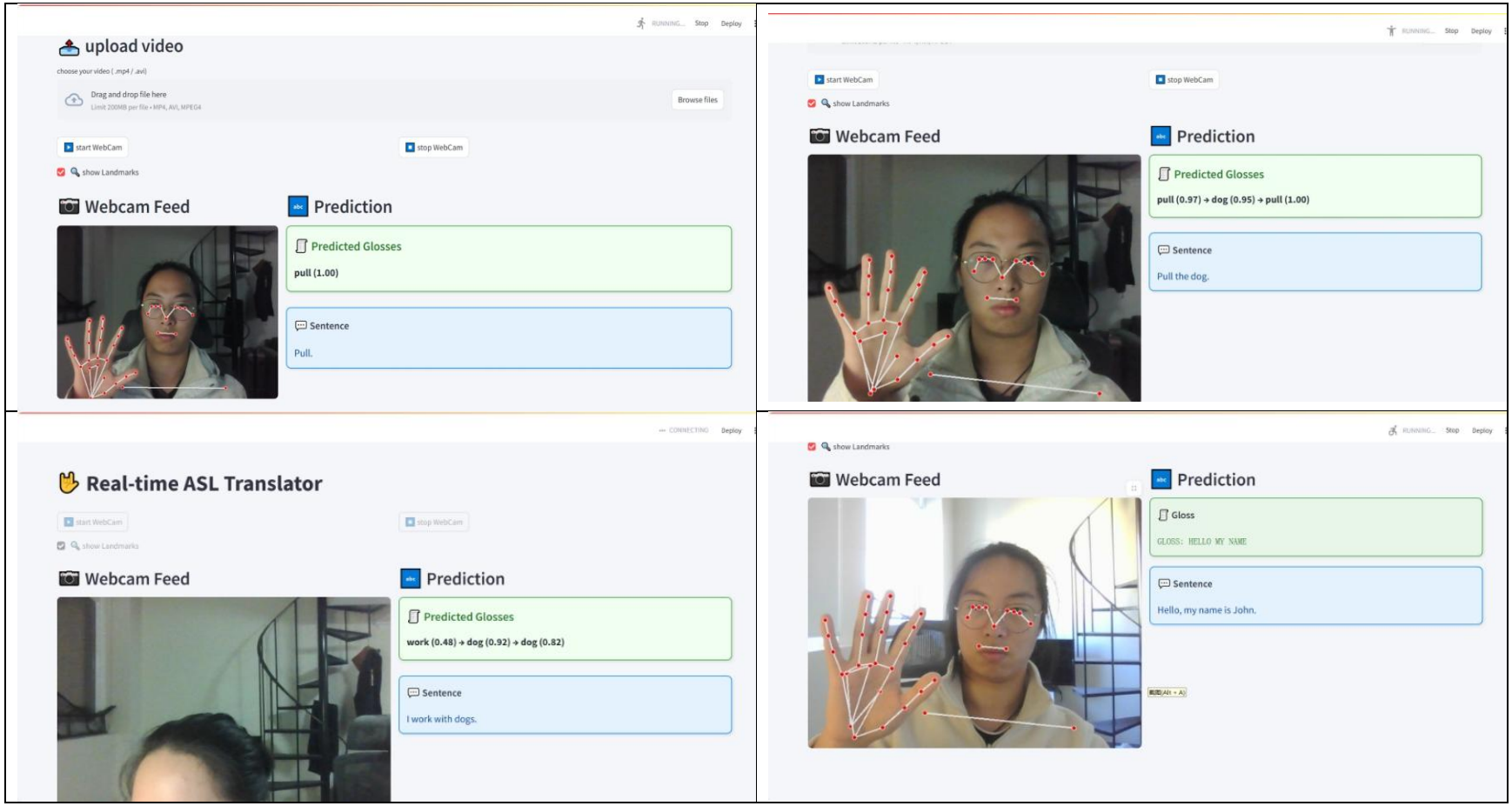
Requirement	User Story	Importance	Issue/Task	Release Version	Comment
Must translate Real-time continuous American Sign Language (ASL) sequences into English text	As a deaf individual, I want my ASL signing to be translated into English text in real-time, so I can communicate with non-signers in virtual meetings.	HIGH	SCRUM-125	V0.1	
	As a non-signer, I want the deaf user's full signing sequence to be translated into clear, correct English sentences, so I can understand their message completely and respond appropriately.	HIGH	SCRUM-146	V0.2	
Must extract Real-time landmarks.	As a user, I would like the system to automatically recognize the sign I'm performing.	HIGH	SCRUM-103	V0.1	

Must enable *webcam input and text output.*	As a user, I want to be able to use the product conveniently at all times, and the inputs and outputs should be easy and understandable.	HIGH	SCRUM-144	V0.2	
Upload a pre-recorded video for translation	As an educator, I want to upload a pre-recorded ASL video and receive an English text translation, so I can share lecture content with non-signing students.	HIGH	SCRUM-143	V0.2	

User interaction

The user interacts with the system through a web application designed for accessibility and ease of use:

- **Accessing the System**: Users navigate to the web app via a browser (e.g., Chrome, Firefox). The homepage features a clean interface with a "Start Translation" button and options to either upload a video or use a webcam.
- **Input Options**:
  - **Live Webcam**: Users grant webcam access, position themselves within the frame, and begin signing. The system processes the video stream in real-time.
  - **Video Upload**: Users drag-and-drop or browse to upload a pre-recorded video (supported formats: MP4, AVI). A progress bar shows processing status.
- **Output display**: Translated English text appears in a dedicated panel below the video, updating in real-time for webcam input or as a complete transcript for uploaded videos. Each sentence includes a score (e.g., 92%) to indicate reliability.
- **Feedback Mechanism (Optional)**: Users can rate translations (e.g., thumbs-up/down) and provide comments via a feedback form, enabling iterative improvements.



Open Questions

Question	Answer	Date Answered
What are the minimum hardware requirements for real-time performance?	Benchmark on a range of devices; likely an 8-core CPU, 16GB RAM, and integrated GPU (e.g., Intel UHD Graphics).	

What fallback platform will be used if Google Meet integration fails?	Zoom will be the fallback, leveraging its API for similar extension deployment.	
How will user privacy be protected with video data?	Process data locally and avoid storing videos without consent.	

Out of Scope

1. Text-to-ASL: Translation Generating ASL animations from English text is not feasible within the project timeline.
2. Speech-to-ASL conversion: Similarly, the current system only converts sign language to text. Converting spoken language into sign language through animation will be explored in later version
3. Multilingual Sign Language Support: Current system only support American Sign Language (ASL). Expansion to other sign languages (e.g., British Sign Language) is planned for future releases.
4. Mobile App Development: The current focus is on a browser extension, not a standalone mobile app.
5. Audio Output: Converting translated text to speech is a future enhancement.
6. Integration with video conferencing Platforms: Future version will include direct integration with Zoom, Google Meet, and other platforms.

Reference materials

Barnett, S., McKee, M., Smith, S. R., & Pearson, T. A. (2011). Deaf sign language users, health inequities, and public health: opportunity for social justice. *Preventing Chronic Disease*, *8*(2), A45–A45.

Dice, D., & Kogan, A. (2021). *Optimizing Inference Performance of Transformers on CPUs*. https://doi.org/10.48550/arxiv.2102.06621

Dill, S., Rösch, A., Rohr, M., Güney, G., De Witte, L., Schwartz, E. & Hoog Antink, C. (2023). Accuracy Evaluation of 3D Pose Estimation with MediaPipe Pose for Physical Exercises. *Current Directions in Biomedical Engineering*, *9*(1), 563-566. https://doi.org/10.1515/cdbme-2023-1141

Li, D., Opazo, C. R., Yu, X., & Li, H. (2020). Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1448–1458. https://doi.org/10.1109/WACV45572.2020.9093512

United Nations. (n.d.). *Reduce inequality within and among countries - United Nations Sustainable Development*. United Nations Sustainable Development. https://www.un.org/sustainabledevelopment/inequality

National Geographic. (2024). *Sign Language*. https://education.nationalgeographic.org/resource/sign-language

World Health Organization. (2021). World Report on Hearing. https://www.who.int/publications/i/item/9789240020481