LLM-Driven Multimodal Video-Text Fusion for Isolated Sign Language Recognition

Sergio Esteban-Romero*
Universidad Politécnica de Madrid
Madrid, Madrid, Spain
sergio.estebanro@upm.es

Cristina Luna-Jiménez
Chair for Human-Centered Artificial Intelligence
Augsburg University
Augsburg, Augsburg, Germany
cristina.lunaj@ugr.es

Manuel Gil-Martín
Universidad Politécnica de Madrid
Madrid, Madrid, Spain
manuel.gilmartin@upm.es

Fernando Fernández-Martínez
Electronics Engineering Departament
Universidad Politécnica de Madrid
Madrid, Madrid, Spain
fernando.fernandezm@upm.es

Elisabeth Andre
University of Augsburg
Augsburg, Germany
andre@informatik.uni-augsburg.de

Abstract

Sign languages are the primary means of communication for deaf communities, but the development of effective automatic recognition systems remains a significant challenge. In this work, we focus on the task of Isolated Sign Language Recognition (ISLR) using a multimodal approach grounded in a Large Language Model (LLM) architecture. We merge modalities, including visual characteristics into the linguistic representation space of LLMs, and perform ablation studies to evaluate the individual contributions of each visual modality to the recognition performance. Experiments are conducted on the AVASAG100 dataset, where our method achieves a weighted F1-score (W-F1) of 70.36±3.00 and a macro F1-score (M-F1) of 62.34±3.18 projecting landmarks extracted from the pose into the LLM's emebdding-space. These results underscore the value of multimodal integration in ISLR and provide guidelines for future research directions.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Human-centered computing** → *Accessibility*.

Keywords

Sign Language Recognition, Human-computer interaction, Accessibility, Multimodal Large Language Models

ACM Reference Format:
Sergio Esteban-Romero, Cristina Luna-Jiménez, Manuel Gil-Martín, Fernando Fernández-Martínez, and Elisabeth Andre. 2025. LLM-Driven Multimodal Video-Text Fusion for Isolated Sign Language Recognition. In *ACM International Conference on Intelligent Virtual Agents (IVA Adjunct '25), September 16–19, 2025, Berlin, Germany.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3742886.3756724

*Correspondence author.

IVA Adjunct '25, Berlin, Germany

ACM ISBN 979-8-4007-1996-7/25/09
https://doi.org/10.1145/3742886.3756724

1 Introduction

Advancements in the field of sign language recognition have focused mainly on two research directions, called Continuous Sign Language Translation (CSLT) [1, 14] and Isolated Sign Language Recognition (ISLR) [5, 24]. Unlike CSLT, which aims to translate continuous video sequences, ISLR aims to accurately predict glosses from independent or isolated video segments. The concept of gloss refers to a written representation used to describe, document or transcribe signs in a spoken language. It facilitates the linguistic study of sign languages and supports their automatization. In ISLR, glosses serve as the output classes predicted by trained classifiers. The input data, however, can vary and may include features extracted from visual encoders, raw images, or landmarks coordinates. In this work, we employ both landmark features and features extracted using Visual Transformers (ViT) to feed Multimodal Large Language Models (MLLM) for performing ISLR. More specifically, the main contributions of the article are as follows:

- We propose a set of state-of-the-art multimodal large language models (MLLM) for recognizing 100 glosses in the AVASAG [4] dataset. The model effectively integrates both linguistic (textual annotations) and visual information (sign language cues) enabling a unified representation that captures both spatial and temporal patterns. Additionally, by leveraging the inherent capacity of Large Language Models (LLMs) to model long-term dependencies, our approach advances the recognition of complex sign language gloss structures.
- We conduct ablation studies to quantify the individual contributions of different visual modalities, providing insights into how each modality affects recognition performance.
- We evaluate the impact of different visual features by comparing landmark-based (MediaPipe) and image-based (Visual Transformers, ViT) representations. These features are projected into the latent space of the Qwen3 model, translating visual input into textual tokens. Furthermore, we investigate

various data pre-processing strategies, including padding and masking, to optimize model performance.
- We identify and analyze systematic misclassifications among visually similar glosses, highlighting key challenges and areas for future improvement in ISLR.

The remainder of the paper is structured as follows. Section 2 reviews related work, covering ISRL and existing MLLM. Section 3 outlines the methodology, beginning with the description of the dataset and the pre-processing steps, followed by details on the extraction of landmarks and features and the proposed MLLM approach. Section 4 presents the experimental setup and key findings. Finally, Section 6 concludes the paper with a summary of the main findings and potential directions for future research.

2 Related Works

This section reviews the literature of the current developments in ISLR and MLLM, emphasizing those approaches that employ LLMs and pose landmarks extracted from videos.

2.1 Isolated Sign Language Recognition

From a technical standpoint, prior work in Isolated Sign Language Recognition from video can be broadly categorized into two main types based on the input features used: 1) RGB-based and 2) pose-based approaches. RGB-based methods process entire video frames and typically with convolutional neural networks (CNNs) to capture spatio-temporal patterns [2, 18]. Although these methods often achieve high accuracy, they are computationally intensive. In contrast, pose-based approaches rely on structured and lightweight representations, using key landmarks from the body and hands, through specialized pose estimation frameworks.

In SignBERT [13], Hu et al. achieved state of the art in multiple SLR benchmarks, treating each hand pose as a visual token and performing one self-supervised masking pre-training stage as in BERT [9], where the model must learn to reconstruct masked tokens. In particular, they only used hand landmarks in their framework, encoding spatial information using spectral-based graph convolutional networks (GCN) [7] and positional embeddings to encode temporal relationships.

Z. Li et al. [19] also proposed a landmarks-based approach which employs three-layer GCN for each of their pose encoders, receiving the landmarks of the face, hands, or body, respectively. In addition to the landmark-based approach, they also incorporate RGB images of the hands through an additional visual encoder. However, the performance remained statistically comparable to the landmark-only version, highlighting the effectiveness of using keypoints and achieving state-of-the-art results across multiple datasets.

Additionally, other pose-based approaches employing variations of transformer architectures have been successfully employed using pose-landmarks for ISLR [5, 23, 24]. For example, the SPOTER architecture [5] implements a query-class at the input of the decoder to reduce the temporal dimensionality of the embeddings obtained at the output of the encoder. Similarly, the Interpretable Transformers [24] version includes an additional weights layer at the input of the encoder to weight the contribution of each landmark.

2.2 Multimodal Large Language Models

The use of MLLMs in Sign Language has recently gained increasing interest since it has been employed in Sign Language Translation problems, achieving significant improvements with respect to other traditional approaches [12, 13, 15, 21, 28].

Gong et al. [12] propose a vector-quantized (VQ) sign language module to obtain a codebook of discrete token representations, which are projected into the semantic space of a LLM to generate the final translated sentence. This approach is based on the assumption that spoken languages are inherently discrete. The model is trained in an autoregressive manner, predicting the next token at each iteration given the video sequence. The codebook produced by the VQ module encodes signs and is aligned with the LLM's semantic understanding of textual words using the Maximum Mean Difference (MMD) [27], which helps bridging the gap between the distributions of visual and textual representations.

Kim et al. [15] conduct an analysis demonstrating that image-based vision language models outperform video-based counterparts in generating descriptions of sign articulation performed by signers in a zero-shot setting. Moreover, they propose a pretraining stage in which RGB images are combined with their corresponding textual descriptions to obtain multimodal features. These features are then aligned with the spoken content in the video, with the objective of performing sign language translation. This alignment is performed through contrastive learning in the same way as proposed by CLIP [26].

In the SCOPE framework proposed by Liu et al. [21], they first encode the landmark motion and align it with the semantic understanding of LLM to identify which glosses appear in a given video performing sign language recognition. Afterwards, a LLM is fine-tuned using Q-LoRA [8] given the glosses and some additional context related to the conversation generating the predicted sentence.

The effectiveness of Sign Language Understanding (SLU) has also proven successful through the architectures validated in the Uni-Sign model [20]. In UniSign [19], mutiple pose-based specialized encoders are employed to extract a variety of features, which are subsequently concatenated and processed by a Spatio-Temporal Graph Convolutional Networks (ST-GCN) [29] to capture short-term dependencies. These representations are then passed through a temporal encoder that integrates the temporal dimension via mean pooling, followed by the concatenation of the resulting features. The concatenated features are projected into the dimensional space of a pre-trained Large Language Model (LM) and fed into it, enabling the fusion of both visual and linguistic modalities. By incorporating landmark features as input tokens, with prior alignment between hand landmarks and their corresponding image features, the model achieves state-of-the-art performance on well-known ISLR benchmarks (e.g. WLSASL100, WLASL2000, etc.).

3 Methodology

This section outlines the methodology used to address the key research questions of this study. We investigate:

(1) Can MLLMs effectively integrate visual and linguistic features for ISLR on the AVASAG100 dataset?

(2) How do different visual inputs—such as pose-based landmarks and ViT-based image features—affect recognition performance?
(3) What is the impact of pre-processing strategies like landmark filtering or the usage of attention masks on gloss prediction accuracy?
(4) What are the main sources of systematic errors, particularly those related to visual similarity between glosses?

The subsections that follow describe the dataset, feature extraction methods, model architectures, and experimental setup used to explore these questions.

3.1 Dataset

The dataset used in this work is AVASAG [4], which consists of a collection of videos featuring sentences related to transportation recorded and signed in German Sign Language (DGS), annotated with glosses and their respective German translation. For our experiments, we focused exclusively on the gloss annotations to perform ISLR. As an initial pre-processing step, we extracted all annotated glosses from the sentence-level annotations. In order to preserve information about the transitions between glosses, each gloss was segmented according to a convention in which the start time of a gloss is defined as the end time of the preceding gloss, applying this rule recursively throughout each sentence sequence [16]. This segmentation approach, illustrated in Figure 1, uses the NOVA platform[1] to ensure that temporal boundaries between glosses are maintained for downstream analysis.

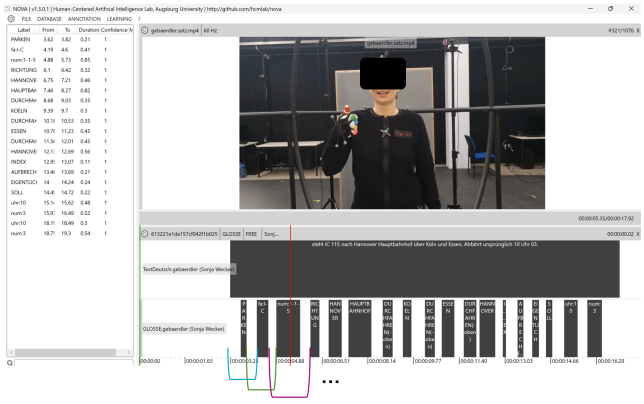

Figure 1: Sample of AVASAG with video annotated with DGS-glosses and German text in NOVA [3].

Following this approach and the structure proposed in other well-known sign language datasets, such as the WLASL dataset [17], we generated a subset of the 100 most frequent glosses in our dataset, which yielded a total set of 4,336 samples with a frequency of gloss repetitions ranging from 470, for the most frequent gloss, to just 14 samples, for the least frequent gloss ($\mu = 43.35$; $\sigma = 54.19$). Subsequently, we split the dataset into train, val, and test subsets with a percentage of 60, 20, 20, respectively, while keeping the class distribution consistent across all sets. In summary, the dataset contained 2,596 samples for training, 849 for validation and 890 for testing. In the remainder of the work, we will refer to this dataset as AVASAG100.

3.2 Feature Extraction

3.2.1 Visual Features Extraction. We employ a ViT[2] to extract visual features from individual frames of sign language videos. The ViT operates by segmenting each frame into fixed-size patches, embedding these patches, and processing the resulting sequence through multiple transformer layers. From the final hidden layer, we explore two widely adopted strategies for feature extraction:

- *CLS Token:* A dedicated, learnable vector that aggregates global information accross the entire input sequence, serving as a holistic representation of the frame.
- *Average pooling:* The mean of all token embeddings from the last hidden state, providing a summary representation that captures the distributed spatial content of the frame.

The resulting embeddings serve as compact representations for each video frame and are subsequently used as input features for downstream classification tasks.

3.2.2 Landmarks Extraction. Compared to image-based features, skeleton-based representations (or landmarks) provide a concise and efficient means of providing models with key motion and posture information. In this work, we used the MediaPipe library [22] to extract 21 landmarks per hand (x, y, z) using the MediaPipe Hands module, and 33 body keypoints using the MediaPipe Pose module, resulting in a total of 75 landmarks per frame.

3.3 Multi-modal Language Models Approach

To integrate the extracted visual features from video frames, we employ MLLMs. MLLMs are particularly effective at modeling long-term dependencies, leveraging their attention mechanisms to capture complex relationships across modalities. This capacity enables the seamless fusion of visual and linguistic information, surpassing the performance achievable by individually processing each modality. Specifically, recent works have demonstrated the state of the art performance of this approach in domains such as human perception analysis [10] and, even more closely aligned to this work, sign language translation and recognition tasks [20].

Our experimental architecture, depicted in Figure 2, is built around the Qwen3 model[3] [11] in its 0.6B parameter configuration. Qwen3 was selected for its extensive multilingual pretraining, covering 119 languages and dialects. To align the different visual modalities with the LLM's semantic space, each input modality is first projected through a dedicated linear layer, which is optimized during training.

3.3.1 Prompt processing. The prompt serves to inject general knowledge, facilitating model alignment and enabling a faster adaptation to the target task. The prompt is first tokenized and then mapped to embeddings using the LLM's pretrained internal embedding layers, ensuring that the prompt information is represented within

[1]NOVA is an open-source annotation tool for multimodal data, commonly used for sign language corpus projects. See: https://github.com/hcmlab/nova

[2]https://huggingface.co/google/vit-base-patch16-384
[3]https://huggingface.co/Qwen/Qwen3-0.6B

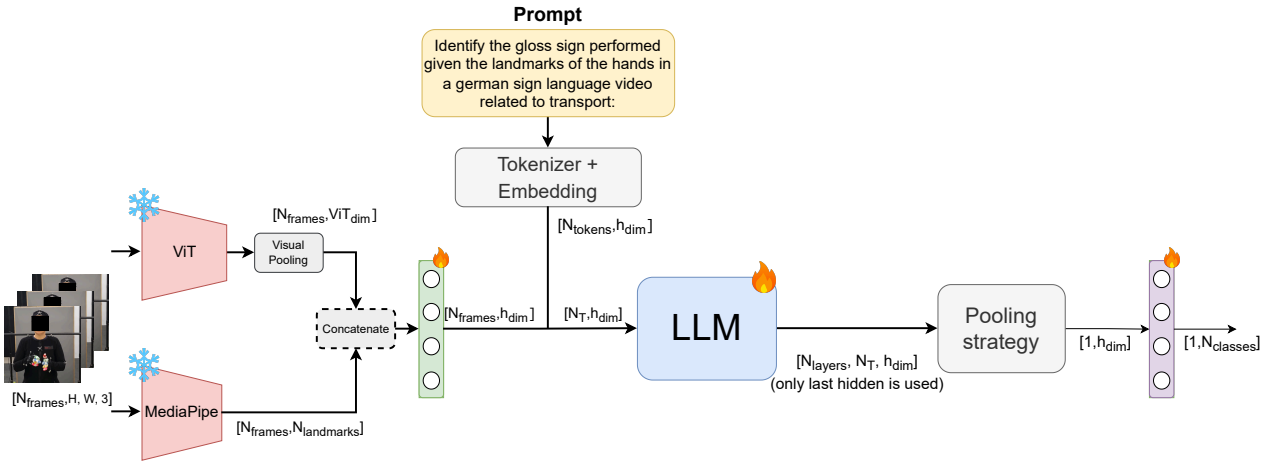

Figure 2: The architecture used to process text, landmarks and images utilizing a LLM. Firstly, the textual prompt describing the task is tokenized. Then, the visual encoder extract ViT features or landmarks and they are projected with a linear layer into the latent space of the LLM (h_{dim}). Later, these visual features are concatenated with the input prompt. After being processed by the LLM, a pooling strategy is applied over the last hidden state of the output obtained. Finally, a linear layer returns $N_{classes}$ logits. The dashed lines indicate components that are only used when using both inpu features. The flame indicates components fine-tuned whereas the snowflake indicates those that remain frozen.

the model's semantic space for its effective integration with other modalities.

3.3.2 Visual encoder. This work investigates how visual information can be effectively integrated with a LLM to advance multimodal sign language understanding and support future research in this area. To this end, two different visual backbone architectures are analyzed:

- *Image encoder (ViT):* We evaluate both the CLS token and the mean of the last hidden state as pooling strategies ('Visual pooling' in Figure 2). The CLS token, commonly employed for classification, provides a global representation of the input, though it may not always be the most informative for every frame. In contrast, average pooling aggregates information across all tokens, yielding a smoothed representation of the frame's spatial content.
- *Landmark encoder*: Landmarks are extracted using the Mediapipe library [22] and flattened before being projected through a linear layer to match the LLM's hidden input dimension (h_{dim} in Figure 2). This projection layer is always fine-tuned, enabling the model to learn an embedding space that aligns landmark information with the LLM's pretrained semantic representations. An important limitation in our dataset is that people performing glosses wear gloves, which occasionally prevents MediaPipe from detecting hand landmarks in certain frames. This is especially problematic for distinguishing glosses which are similar except for subtle fingers movements. To address this, we conducted ablation experiments to study whether removing finger landmarks impacts overall performance.

After extracting and processing the features, visual representations are projected using a linear layer to h_{dim} and combined with the prompt tokens to form the model input.

In our approach, we process all frames in the input videos, which significantly reduces the dimensionality of visual data while preserving essential motion and pose information. By encoding each frame as a distinct token based on its landmark configuration, we are able to treat the entire video as a temporal sequence suitable to be processed by the LLM. This strategy enables us to utilize the full temporal resolution of the video but the total number of frames that can be processed is ultimately constrained by either the available memory or the maximum input sequence length of the model.

The combined prompt and visual tokens are then processed by the LLM, which is always fine-tuned for our tasks. Since the LLM is pre-trained exclusively on text, fine-tuning is necessary to adapt it to the new multimodal domain. In this regard, it is important to remark that freezing the LLM led to suboptimal results in preliminary experiments.

Extracting meaningful representations from LLMs for classification is not trivial, as the most informative features do not always correspond to the last encoder output [6, 25]. Following recent findings, we apply mean-pooling over the last hidden state, assigning equal importance to each token's hidden representation to preserve global temporal information. The pooled embedding is processed by a final linear layer, which is also fine-tuned, to produce logits for each of the 100 gloss classes ($N_{classes}$) in our problem. The gloss with the highest probability is finally selected as the predicted class for each video.

4 Results

This section provides a general overview of the experiments conducted in this work with a detailed analysis and comparison between the reported results.

4.1 Experimental Conditions

In this section, we report the experimental setup employed to perform the experiments and describe the hyperparameters used to train the models and the metrics used to evaluate them.

4.1.1 Hyper-parameters of the models. The ablation experiment studies were performed using Bidirectional Long Short Term Memory (Bi-LSTMs) models with 2 layers and a hidden dimension of size 256 with a linear layer on top to perform the final classification. These baseline models were trained with a learning rate of 10^{-5} for a maximum of 500 epochs. The experiments using LLM were performed with a learning rate of 10^{-6} for a maximum of 100 epochs. In both training scenarios, we use the Adam optimizer, batch size of 16, an early stop of 15 epochs without improvement in validation loss and CrossEntropyLoss as the cost function to measure the models performance. The Graphics Processing Unit (GPU) utilized is a NVIDIA GeForce RTX3090 24GB.

4.1.2 Metrics. Since in the dataset there is a notable imbalance in the number of certain glosses, we employed the Weighted-F1 (W-F1) and Macro-F1 (M-F1) metrics. W-F1 computes the weighted mean by considering the number of samples in each class and the priors. In contrast, M-F1 treats every class as equal, regardless of the amount of samples or whether they are unbalanced by calculating the arithmetic mean of all F1 scores for each class. In terms of mathematics, they are described as follows:

$$\text{F1} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FP} + \text{FN}}$$

$$\text{W-F1} = \sum_{i=1}^{N} w_i * F1_i \; ; \text{M-F1} = \frac{\sum_{i=1}^{N} F1_i}{N}$$

where N represents the total number of glosses and w_i is the percentage of samples per glosse. FP, TP, and FN represent the number of False-Positives, True-Positives, and False-Negatives, respectively.

4.2 Analysis of Results

In this section, we present an ablation study to identify the individual contribution of each visual modality and also the influence of different preprocessing methods that have been applied, such as filtering subsets of landmarks with a high number of missing values or the usage of attention mask in the LLM for those padded sequences. Moreover, an error analysis is presented to determine the weaknesses of our contribution so they can be properly addressed in future studies.

4.2.1 Comparison of Input Features. In this section, we compare both visual features (ViT features obtained from RGB-images vs. landmarks) to measure how the model performs when they are individually provided as input. To evaluate the advantages of our approach, we compare it against a Bi-LSTM baseline commonly used for modeling temporal dependencies in sequences. Table 1 reports the results obtained using Bi-LSTM on the analyzed features. MediaPipe landmarks consistently outperform all RGB-based features. When the LLM is fed with the ViT features extracted directly from the RGB-images, the performance is poor compared to the landmarks-based approach, although it still surpasses the Zero Rule (W-F1 = 2.02 ± 0.92, M-F1 = 0.19 ± 0.29) achieving an increase of ΔW-F1=+3, 19% and ΔM-F1=+1, 41% in percentage points.

Model	Visual Feature Encoder	Visual Pooling	W-F1↑ ±CI	M-F1↑ ±CI
Bi-LSTM	vit-base-patch16-384	CLS Token	3.78±1.25	0.88±0.61
Bi-LSTM	vit-base-patch16-384	Avg. Pooling	3.38±1.19	0.76±0.57
Bi-LSTM	MediaPipe	✗	19.7±2.61	9.10±1.89
LLM	vit-base-patch16-384	CLS token	5.03±1.44	1.50±0.79
LLM	vit-base-patch16-384	Avg. Pooling	5.21±1.46	1.60±0.82
LLM	MediaPipe	✗	55.14±3.26	46.52±3.27

Table 1: Baseline results obtained for each visual modality using a Bi-LSTM and LLMs. CI stands for 95% Confidence Interval.

To understand this behavior, we computed the pairwise cosine similarity matrix per video comparing the relationship between the embeddings extracted from each frame of the video. In Figure 3, we displayed a sample of the confusion matrix for one of the videos. We observed that the majority of similarity values are close to one, indicating highly consistent frame representations.

To objectively quantify this observation, we calculate the average of all pairwise similarities for each video and then report the overall mean and standard deviation throughout the dataset. We surprisingly find $\mu = 0.9999$ and $\sigma \approx 0$, suggesting that the frame embeddings within each video are nearly identical on average, not providing genuine information across embeddings of different timesteps.

This behavior may be explained by the fact that the pre-trained checkpoints of ViT are proficient in providing general image descriptors. However, in the case of sign language, where the global context of the image barely varies between contiguous frames (e.g. same background, same person signing), ViT may not be sufficient to capture the actual semantic encoded in the subtle and rapid variations of facial expressions and hand movements that takes place while performing a sign. This is consistent with the existing literature as reported by the current state-of-the-art model in numerous sign language benchmarks, such as Uni-Sign [19], which showcases a small performance improvement when incorporating features of the hand image considering that they only process the hands via prior cropping with general-purpose pre-trained image encoders. Therefore, either downstream adaptation of the ViT arises as necessary, or the usage of Vision Language (VL) models to enforce the ViT model to focus on hands via textual conditioning.

However, when using landmarks in the exact same configuration, it can be seen how model performance increases significantly in ΔW-F1=+49.93% and ΔM-F1=+44.92% percentage points compared to use ViT features with average pooling. These findings further support the use of landmarks as input features for our analysis and highlights the suitability of combining visual features as contextual representation for the LLM.

4.2.2 Ablation Studies of Masking Strategies and Filtering. One of the main problems we encounter when using MediaPipe landmarks is the scarcity of hand landmarks in some videos, which may mislead the model in its learning process. Therefore, we explore their contribution to the final model by directly removing them, creating a subset containing only pose landmarks. The filtering of such a priori relevant information is supported by having also pose features that incorporate hand landmarks, although with significant fewer descriptors than those provided by the dedicated hand landmark detector.

When working with a LLM and batched input sequences derived from videos of varying duration, padding is necessary to standardize sequence lengths to the longest in the batch. Consequently, to ensure that the model focuses only on tokens that contain meaningful information, we also investigated the impact of incorporating attention masks.

In Table 2 the reported results show that removing the subset corresponding to the hand landmarks provided by MediaPipe significantly improves the model performance when no attention masks are applied (ΔW-F1=+19.19% and ΔM-F1=+21.36% percentage points) and also when padding is masked (ΔW-F1=+15.26% and ΔM-F1=+15.82%). These results indicate that the hand landmarks provided by MediaPipe are not a good descriptor for this specific problem, so different hand landmark and feature extractors will be explored in future works.

Regarding the use of attention masks, performance improves although it is not statistically significant when the MediaPipe hand landmarks are not filtered (ΔW-F1=+4.18% and ΔM-F1=+4.01%). The same behavior is noticed when hand landmarks are filtered (ΔW-F1=+0.21% and ΔM-F1=−1.53%), indicating the need for further investigation. However, given that our dataset comprises relatively short video sequences, we hypothesize that the benefits of attention masking would become more pronounced with longer sequences.

Despite the poor performance achieved when using ViT features, we explore the possibility of integrating information from both visual modalities to explore if it is possible to enrich each frame representation. Therefore, we concatenate both inputs obtaining a multimodal embedding that is projected into the semantic space of the LLM using a linear layer. In particular, as there is no statistically significant difference, we use the embeddings obtained from ViT taking the CLS token. The results obtained using the CLS token ViT embeddings and the filtering the MediaPipe landmarks using attention masks are W-F1=53.22±3.27 and M-F1=42.72±3.25. These results do not outperform those obtained using only the landmarks visual modality. This may be due to the inclusion of the ViT branch, which could hinder overall performance by introducing more confusion than useful information. It would be worth to optimize the image branch and reattempt the multimodal integration approach.

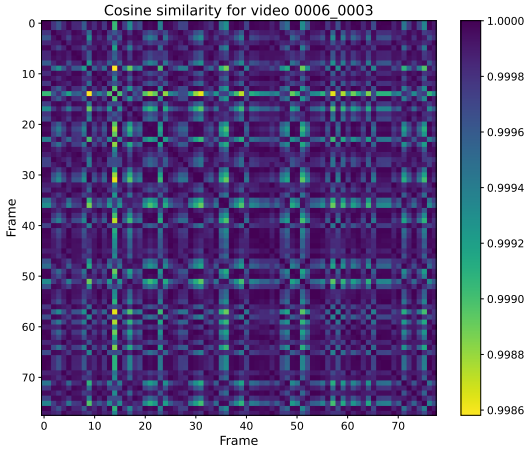

Figure 3: Cosine similarity across the embeddings representing each frame in the video with lowest global mean cosine similarity

Use Attention Masks	Filter MediaPipe Landmarks	W-F1↑ ±CI	M-F1↑ ±CI
✗	✗	50.96±3.28	42.51±3.24
✓	✗	55.14±3.26	46.52±3.27
✗	✓	70.15±3.01	63.87±3.16
✓	✓	70.36±3.00	62.34±3.18

Table 2: Ablation Study of the contribution of Attention Masks and filtering on MediaPipe landmarks using the LLM.

4.2.3 Error Analysis. In this section, we provide an analysis of the most relevant concerns that we discovered through our experimentation. Firstly, we discovered that ViT features lack enough variability to individually solve the task or to complement other features such as the landmarks. Although we quantitatively measured this limitation, Figure 3 presents the cosine similarity matrix of the embeddings corresponding to the video with the highest standard deviation, computed over the mean pooled outputs from the last hidden state of the Vision Transformer (ViT). It should be noted that the largest differences observed in cosine similarity are on the order of $\Delta = 1.4 \times 10^{-3}$, indicating only minimal variation between the compared embeddings. These results highlight that features extracted from a pre-trained ViT model without adaptation to the task, as used in this study, fail to capture nuanced differences between frames, suggesting that alternative architectures or task-specific fine-tuning may be necessary directions for future research.

Figure 4 represents the confusion matrix for the predictions associated with our best result from Table 2 where the hand landmarks are filtered and an attention mask is applied so the model does not take into account padding. Although many classes are classified with high precision, as indicated by the prominent diagonal values, systematic misclassifications are also evident. For example, glosses with similarities in how they are articulated, such as

variants of 'num' (e.g., 'num.1', 'num.ord.1', etc.) exhibit significant misclassifications among each other, suggesting limitations in the discriminative capacity of the model for fine-grained numerical entities. Even more notable but also related, it is the issue that arises with glosses related to the finger alphabet (e.g. 'fa.a', 'fa.b', 'fa.c' and 'fa.d'). These errors may result from overlapping features, especially due to filtering of hand landmarks, suggesting that the incorporation of hand features may help to mitigate this limitation. Furthermore, the model shows a notable bias towards predicting wrongly the gloss 'index', likely due to being highly overrepresented in the training data.

5 Limitations

In this section, we will highlight the main limitations associated with our research. Regarding the data acquisition step, described in Section 3.1, the segmentation process was carried out manually, which may result in segments that are not ideal to be processed by the model when discerning glosses.

Also related with the dataset, the class imbalance presents a series of challenges as the models exhibit a notable bias in predicting the majority class in the dataset. This can be clearly seen in Figure 4 where the gloss with the highest number of occurrences "index" is wrongly predicted on several occasions. Additionally, the existing similarity between groups of glosses with similar articulation (i.e. numbers or finger alphabet glosses) highlight potential limitations that could be addressed through more fine-grained hand descriptors or by increasing temporal resolution to better distinguish them.

Despite notably improving efficiency and simplifying the feature extraction process when using a ViT to extract features from each frame, the dimensionality reduction techniques applied in this work entail a notable information loss. This could potentially explain the similarity between the frame representations that are obtained for each frame in our dataset as we are obtaining a high-level description of the image that is unable to capture the nuances required to accurately discern between signs (i.e., finger orientations, hand positioning, facial expressions, etc.). Hence, the acquisition of a unique global representation for each frame may limit the capabilities of the model to exploit these complex descriptors. Therefore, a more effective approach might involve cropping the images to focus solely on specific areas such as the hands, arms, or face and using these cropped regions as input to the model.

Additionally, the concatenation of landmarks into a unique flattened representation to be processed as a token may be suboptimal, especially when dealing with missing landmarks. It may be worth exploring to separate landmarks into body structure groups (i.e. hands, pose, face, etc.) and process them separately allowing the model to exploit their individual temporal and spatial dynamics without being constrained to process them simultaneously. Another potential solution may involve attention mechanisms that give weight to each landmark according to their relevance for effectively predicting the expected outcome. This solution would help in mitigating the impact of scenarios where some grops of landmarks are not present or obtained with a high confidence.

6 Conclusions

In this work, we proposed a multimodal approach to perform Isolated Sign Language Recognition (ISLR) that takes advantage of a compact Large Language Model (LLM) with 0.6 billion parameters. Our method was evaluated on a subset of the AVASAG100 dataset, which contains videos of 100 German Sign Language glosses related to the domain of transportation, with each video corresponding to a single gloss instance. Our method achieves notable high performance, with a W-F1 score of 70.36±3.00 and a M-F1 score of 62.34±3.18 when applying landmark filtering, indicating the need for more expressive landmark descriptors. This limitation could potentially be addressed by employing alternative open-source libraries that can overcome MediaPipe limitations. Future research will also focus on obtaining image-based features able to capture better temporal variance and hand information than the employed models, with the scope of seamlessly integrating the visual modality representations before being provided to the LLM. However, the LLM is demonstrated to be capable of integrating features from multiple visual modalities, such as hand and pose landmarks and image frames, by aligning them through fine-tuning, enabling the model to bridge the gap between heterogeneous inputs with minimal computational overhead.

Acknowledgments

This contribution is funded by the German Ministry for Education and Research (BMBF) through the BIGEKO project, grant number 16SV9094. Sergio Esteban-Romero research was supported by the Spanish Ministry of Education (FPI grant PRE2022-105516). This work was funded by Project ASTOUND (101071191 — HORIZON-EIC-2021-PATHFINDERCHALLENGES-01) of the European Commission and by the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C22), TRUST-BOOST (PID2023-150584OB-C21) and BeWord (PID2021-126061OB-C43), funded by MCIN/AEI/ 10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR".

References

[1] Junseok Ahn, Youngjoon Jang, and Joon Son Chung. 2024. Slowfast Network for Continuous Sign Language Recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3920–3924. doi:10.1109/ICASSP48485.2024.10445841

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision*.

[3] Tobias Baur, Alexander Heimerl, Florian Lingenfelser, Johannes Wagner, Michel F. Valstar, Björn Schuller, and Elisabeth André. 2020. eXplainable Cooperative Machine Learning with NOVA. *KI - Künstliche Intelligenz* (19 Jan 2020). doi:10.1007/s13218-020-00632-3

[4] Lucas Bernhard, Fabrizio Nunnari, Amelie Unger, Judith Bauerdiek, Christian Dold, Marcel Hauck, Alexander Stricker, Tobias Baur, Alexander Heimerl, Elisabeth André, Melissa Reinecker, Cristina España Bonet, Yasser Hamidullah, Stephan Busemann, Patrick Gebhard, Corinna Jäger, Sonja Wecker, Yvonne Kossel, Henrik Müller, Kristoffer Waldow, Arnulph Fuhrmann, Martin Misiak, and Dieter Wallach. 2022. Towards Automated Sign Language Production: A Pipeline for Creating Inclusive Virtual Humans. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments* (Corfu, Greece) *(PETRA '22)*. Association for Computing Machinery, New York, NY, USA, 260–268. doi:10.1145/3529190.3529202

[5] Matyáš Boháček and Marek Hrúz. 2022. Sign Pose-Based Transformer for Word-Level Sign Language Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 182–191.

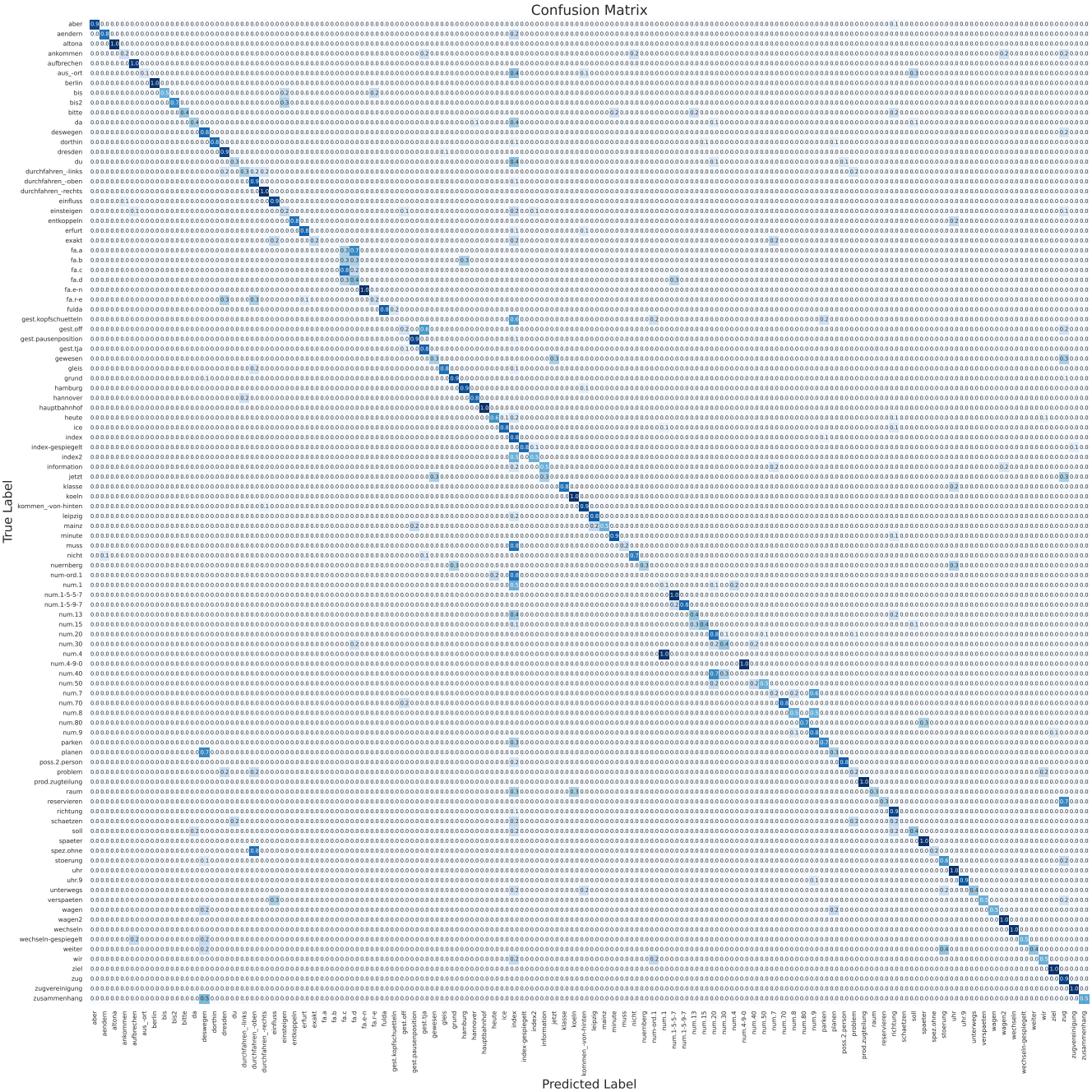

Figure 4: Confusion matrix showing the classification results for gloss recognition. Diagonal elements represent correctly classified glosses, while off-diagonal elements indicate misclassifications.

[6] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. 2025. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181* (2025).

[7] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. 2019. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2272–2281.

[8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314 [cs.LG] https://arxiv.org/abs/2305.14314

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[10] Sergio Esteban-Romero, Iván Martín-Fernández, Manuel Gil-Martín, David Griol-Barres, Zoraida Callejas-Carrión, and Fernando Fernández-Martínez. 2024. LLM-Driven Multimodal Fusion for Human Perception Analysis. In *Proceedings of the 5th on Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor* (Melbourne VIC, Australia) *(MuSe'24)*. Association for Computing Machinery, New York, NY, USA, 45–51. doi:10.1145/3689062.3689084

[11] An Yang et al. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

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

[13] Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021. Signbert: pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11087–11096.

[14] Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. 2023. CoSign: Exploring Co-occurrence Signals in Skeleton-based Continuous Sign Language Recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Paris, France, 20619–20629. doi:10.1109/ICCV51070.2023.01890

[15] Jungeun Kim, Hyeongwoo Jeon, Jongseong Bae, and Ha Young Kim. 2024. Leveraging the Power of MLLMs for Gloss-Free Sign Language Translation. arXiv:2411.16789 [cs.CV] https://arxiv.org/abs/2411.16789

[16] Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. Public DGS Corpus: Annotation Conventions / Öffentliches DGS-Korpus: Annotationskonventionen. doi:10.25592/uhhfdm.10251

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

[18] Dongxu Li, Cristian Rodriguez-Opazo, Xin Yu, and Hongdong Li. 2019. 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)* (2019), 1448–1458. https://api.semanticscholar.org/CorpusID:204851909

[19] Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2025. Uni-Sign: Toward Unified Sign Language Understanding at Scale. *arXiv preprint arXiv:2501.15187* (2025).

[20] Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2025. Uni-Sign: Toward Unified Sign Language Understanding at Scale. *arXiv preprint arXiv:2501.15187* (2025).

[21] Yuqi Liu, Wenqian Zhang, Sihan Ren, Chengyu Huang, Jingyi Yu, and Lan Xu. 2025. SCOPE: Sign Language Contextual Processing with Embedding from LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 5739–5747.

[22] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Perceiving and Processing Reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*. https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf

[23] Cristina Luna-Jiménez, Lennart Eing, Annalena Aicher, Fabrizio Nunnari, and Elisabeth André. 2025. Lightweight Transformers for Isolated Sign Language Recognition. In *Proceedings of the 27th International Conference on Multimodal Interaction* (Canberra, Australia) *(ICMI '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3716553.3750772

[24] Cristina Luna-Jiménez, Manuel Gil-Martín, Ricardo Kleinlein, Rubén San-Segundo, and Fernando Fernández-Martínez. 2023. Interpreting Sign Language Recognition using Transformers and MediaPipe Landmarks. In *Proceedings of the 25th International Conference on Multimodal Interaction* (Paris, France) *(ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 373–377. doi:10.1145/3577190.3614143

[25] Iván Martín-Fernández, Sergio Esteban-Romero, Jaime Bellver-Soler, Fernando Fernández-Martínez, and Manuel Gil-Martín. 2024. Larger Encoders, Smaller Regressors: Exploring Label Dimensionality Reduction and Multimodal Large Language Models as Feature Extractors for Predicting Social Perception. In *Proceedings of the 5th on Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor* (Melbourne VIC, Australia) *(MuSe'24)*. Association for Computing Machinery, New York, NY, USA, 20–27. doi:10.1145/3689062.3689083

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[27] Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. 2016. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems* 29 (2016).

[28] Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation. arXiv:2405.04164 [cs.CV] https://arxiv.org/abs/2405.04164

[29] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (New Orleans, Louisiana, USA) *(AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, Article 912, 9 pages.