# SAGE: Segment-Aware Gloss-Free Encoding for Token-Efficient Sign Language Translation
## *Supplementary Material*

Low Jian He  Ozge Mercanoglu Sincan  Richard Bowden
CVSSP, University of Surrey, United Kingdom
{jianhe.low, o.mercanoglusincan, r.bowden}@surrey.ac.uk

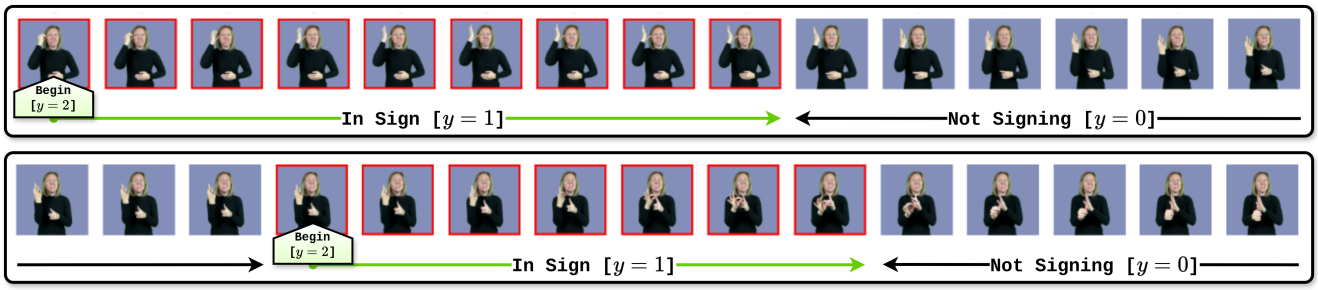


Figure 1. **Sign Segmentation Labeling Scheme.** Given a continuous sign language video, the segmentation model predicts a per-frame label sequence: $y = 2$ (start of a sign), $y = 1$ (continuation of a sign), and $y = 0$ (non-signing intervals or transitional movements).

**Supplementary Material Overview.** Our supplementary material includes additional details and analysis to support the main paper. We first elaborate on the visual tokenization framework, including runtime measurements, example outputs, and further ablations. We also provide qualitative visualizations demonstrating how continuous sign language sequences are segmented into coherent sign-level subclips. For the full model architecture, we include additional alignment map visualizations between visual tokens and pseudo-gloss candidates, as well as example translation outputs to illustrate end-to-end performance.

## A. Additional Details on Visual Tokenization

To perform visual tokenization, we adopt *Hands-On* [2], a segmentation model trained to detect individual sign boundaries in continuous sign video. The model operates on per-frame hand representations $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T\}$ extracted via the HaMeR framework [3], and corresponding 3D body pose sequences $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_T\}$ derived from [1]. These modalities are fused into a multimodal sequence $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_T\}$, which is processed by a Transformer to predict per-frame sign segmentation labels $\mathbf{y} = \{y_1, y_2, \ldots, y_T\}$, where $y_t \in \{0, 1, 2\}$.

The labels $y_t \in \{0, 1, 2\}$ follow a BIO-style scheme, where $y_t = 2$ (**B**) marks the start of a sign, $y_t = 1$ (**I**) indicates continuation within a sign, and $y_t = 0$ (**O**) denotes non-signing or transitional phases (e.g., co-articulation). This enables temporally localized segmentation of continuous signing streams into meaningful visual units (see Fig. 1). To construct visual tokens, we split the video at each $y_t = 2$ (i.e., **B**-tag), treating it as the start of a new subclip. Each segment continues until the next $y_t = 2$, ensuring all predicted signs are extracted. Importantly, we include intermediate $y_t = 0$ (i.e., **O**-tag) frames within a segment, as they often contain co-articulatory motion that preserves linguistic continuity. Our initial experiments showed that discarding these frames reduced translation quality. With this approach, we measured the segmentation runtime across the entire PHOENIX14T dataset and found that generating the segmentation annotations requires only *5 mins 24 secs*, indicating the method's scalability to larger datasets. In addition, we also illustrate how these segments are distributed over continuous signing videos by providing visualizations in Section B.1, where long frame sequences are annotated to show which intervals were selected as segments.

While the segmentation model performs well in general, it was not explicitly trained on our target datasets, which can lead to noisy predictions. For instance, consecutive frames may both be labeled $y_t = 2$, and some signing content may be misclassified as $y_t = 0$, fragmenting valid signs or omitting segments entirely. To address this, we introduce a

| Fallback Segment Length | BLEU4 | ROUGE |
|---|---|---|
| 6 | 22.60 | 47.70 |
| **10** | **23.81** | **49.08** |
| 14 | 22.37 | 47.75 |

Table 1. **Ablation on fallback segment length.** We compare the translation performance when using different fallback lengths.

| | PHOENIX-14T | | |
|---|---|---|---|
| **Split** | **Train** | **Dev** | **Test** |
| Average Number of Frames | 117 | 107 | 101 |
| Average Number of Tokens | 15 | 14 | 13 |
| Reduction Ratio | 0.128 | 0.131 | 0.129 |

Table 2. **Reduction Ratios of PHOENIX14T Dataset Split.** Reported averages of frames and tokens are calculated per video.

lightweight post-processing strategy. First, we remove consecutive **B**-tags by retaining only the first $y_t = 2$ prediction. Second, to ensure full temporal coverage, we partition long stretches of **O**-labeled frames (i.e., $y_t = 0$) into fallback segments of fixed length. These segments act as backup tokens in cases where the segmentor fails to detect signs. We ablate fallback segment lengths of 6, 10, and 14 frames (Table 1). A 10-frame setting yields the best balance between segment granularity and temporal coverage. Shorter segments likely lack sufficient temporal context, while longer segments risk merging multiple distinct gestures.

As described in the main paper, each identified sign segment is represented as a single visual token. Since a segment corresponds to a subclip of consecutive video frames, we compute a token by averaging the visual features over the temporal axis of the segment. This ensures that each token captures the semantics of the entire signing motion while maintaining a fixed dimensionality. This procedure effectively transforms a continuous sign video into a compact sequence of visual tokens, one per segment. To quantify the degree of temporal compression, we compute the *reduction ratio* as the average number of frames per video divided by the average number of visual tokens produced:

$$\text{Reduction Ratio} = \frac{\text{Avg. \# frames per video}}{\text{Avg. \# tokens per video}}.$$

Table 2 reports this ratio across the PHOENIX14T splits. We observe a consistent ratio of approximately $\sim$ 0.13, indicating that our tokenized representation is nearly $8\times$ shorter than the original frame sequence. This improves significantly over recent state-of-the-art methods, which typically achieve a reduction ratio of 0.25. Our approach thus enables greater computational efficiency, making it well-suited for large datasets or resource-constrained hardware.

# B. Additional Qualitative Results

## B.1. Sign Segment Visualization

In Fig. 2, we present qualitative visualizations of how long continuous sign language videos are segmented into subclips based on B-tags (i.e., frames where $y_t = 2$). To aid interpretability, we apply alternating red and blue color bands to indicate segment boundaries, with each color transition denoting the start of a new segment. Above each frame, we annotate the predicted segmentation label $y_t \in \{0, 1, 2\}$, allowing readers to trace the segmentation process frame by frame. As discussed in Section A, our framework incorporates co-articulatory transitions (i.e., frames with $y_t = 0$) within a segment, as long as they occur before the onset of the next B-tag. This is evident in the visualizations, where segments may include the co-articulation of signs.

From a qualitative standpoint, we observe that segment boundaries often align with meaningful linguistic and prosodic cues such as shifts in hand shape or direction of motion. For example, segmentation typically occurs when a signer transitions between distinct hand configurations or executes large-scale movements, such as raising or lowering the hands. While occasional over-segmentation is observed, where a long, continuous sign is divided into multiple segments, this generally occurs in signs involving complex articulatory dynamics. Importantly, these segmentations still preserve semantic structure, as corroborated by our localization and alignment map results in Section B.2.

## B.2. Alignment Map Visualization

In Fig. 3, we present additional qualitative examples of alignment maps that are slightly simplified compared to those shown in the main paper. These visualizations are intended to highlight the approximate localization behavior of our model without explicitly annotating the one-to-one correspondences between specific pseudo-glosses and visual token labels. The alignment maps are generated by computing the dot-product similarity between the visual token sequence and the language embeddings of pseudo-gloss candidates, forming a soft alignment matrix. The horizontal axis represents the visual tokens (one block per segment), while the vertical axis lists the pseudo-gloss candidates derived from the spoken language reference.

To guide interpretation, we display the ground truth gloss annotations above each alignment map. We highlight in green any pseudo-gloss candidates that align with visual tokens whose ground truth gloss labels are either exact matches or strong semantic paraphrases. These highlights allow us to qualitatively verify whether the model's attention is correctly focused. Across all four examples, we observe that whenever a relevant gloss exists among the pseudo-gloss candidates, a clear alignment is formed in the map. This suggests that the model effectively localizes se-

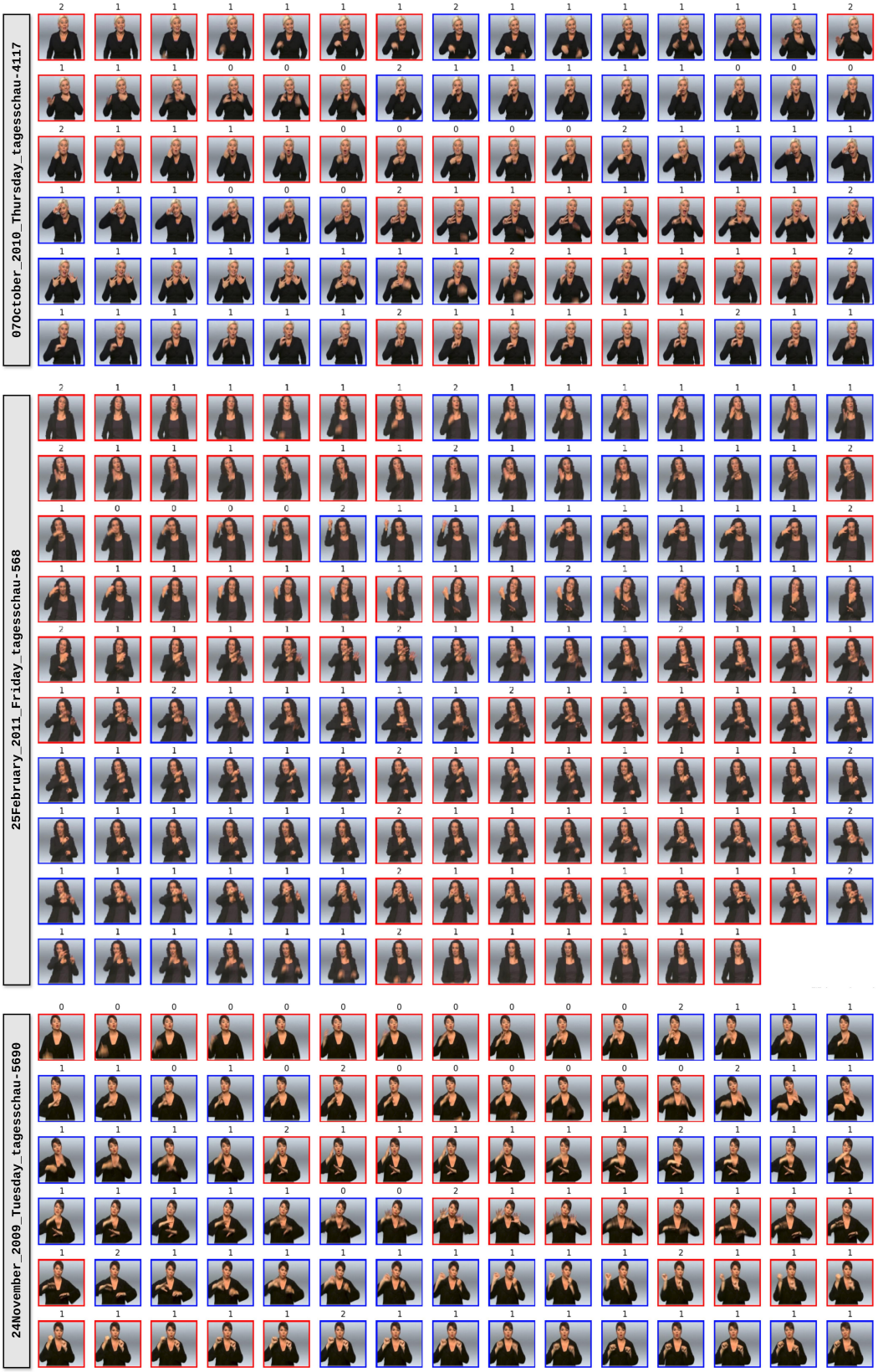


Figure 2. **Qualitative Visualization of Segmented Continuous Sign Videos.** Each sequence illustrates how the segmentation model partitions the video into distinct sign segments. Alternating red and blue bands denote the individual segments, clearly indicating their temporal boundaries and extent within the video stream.

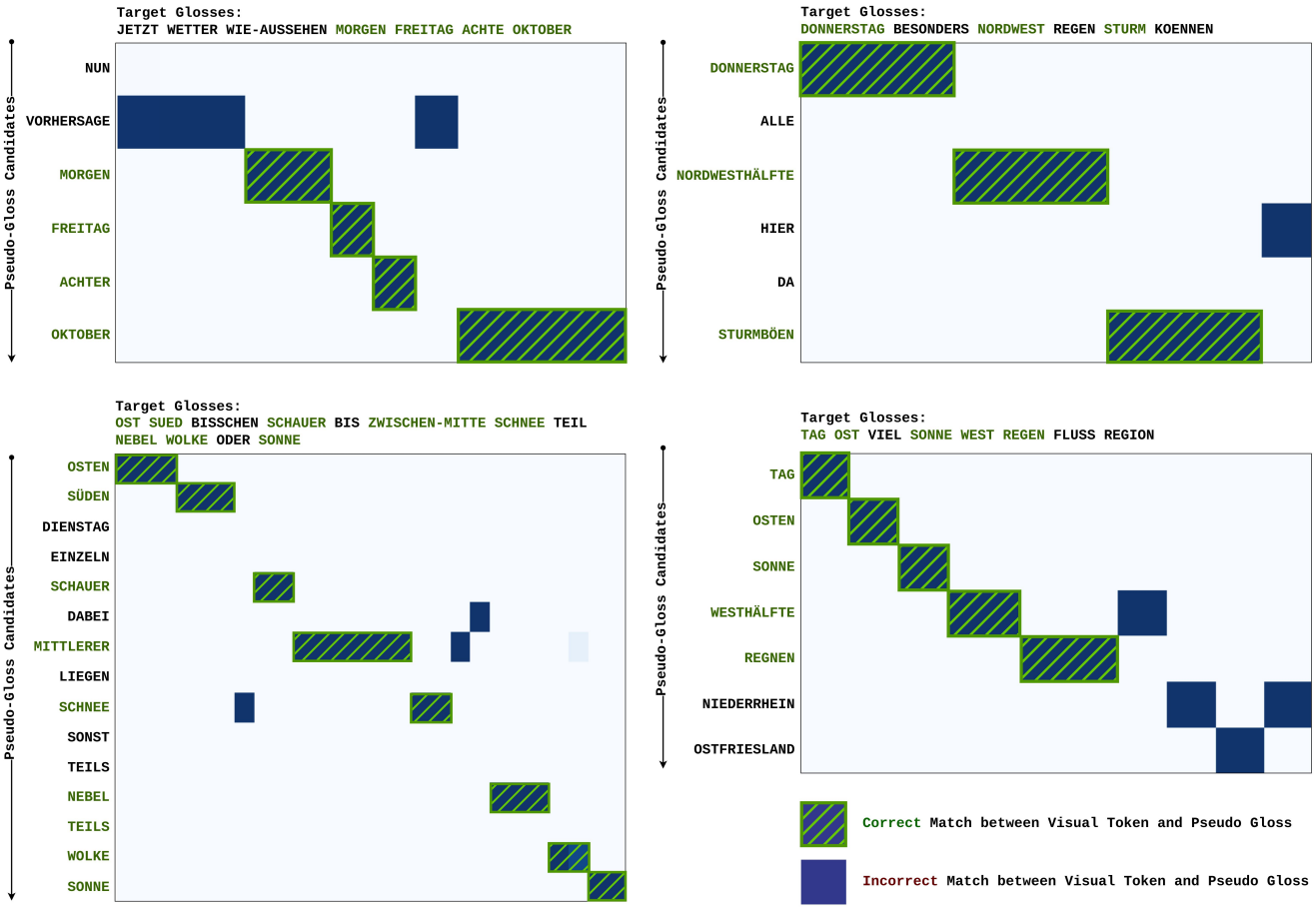


Figure 3. **Qualitative Visualizations of Visual Token Alignment Maps.** We provide additional qualitative alignment maps illustrating the correspondence between visual tokens and pseudo-gloss candidates. Each matrix shows the dot-product similarity between the visual token sequence (horizontal axis) and pseudo-gloss embeddings (vertical axis). Ground truth glosses are shown above each matrix. Visual tokens that match or closely align semantically with the pseudo-glosses are highlighted in green.

mantically meaningful segments even in the absence of explicit gloss supervision. These results further demonstrate the effectiveness of our fine-grained visual-language pretraining, as the visual tokens exhibit consistent alignment with semantically relevant language embeddings.

### B.3. Translation Examples

In Table 3, we present qualitative examples of translation outputs generated by our model, categorized into three groups: (1) Success cases, where the predicted translation matches the reference exactly; (2) Partial success cases, where the translation differs in form but conveys a similar meaning; and (3) Failure cases, where the predicted output diverges significantly from the ground truth. Across the PHOENIX14T test set, we observe that the majority of predictions fall into the partial success category. These translations often retain semantic fidelity to the reference but differ in sentence structure or lexical choice.

Exact matches are comparatively rare, which is not a major concern given that sign glosses typically serve as abstract, content-heavy representations that do not encode full grammatical or functional language. As such, expecting a one-to-one match in the spoken language output is unrealistic. Additionally, we observe that failure cases generally occur when the reference spoken sentence includes information not directly observable in the sign video. For instance, in the final example, the reference states “temperatures drop to...” while the model predicts “temperatures climb to highs of...”. At first glance, this appears to be a clear error. However, the signed gloss sequence for this example is TEMPERATUR DREIZEHN ZWISCHEN FUENF GRAD, which translates roughly to “temperature thirteen between five degrees.” Notably, there is no visual evidence in the signing that indicates whether temperatures are rising or falling. Thus, the model’s confusion stems from the lack of disambiguating cues in the visual input and highlights the broader challenge of inferring contextually implied meaning when it is absent from the signing.

| | | |
|---|---|---|
| **Success Cases** | | |
| **Reference** | ähnliches wetter auch am donnerstag | *similar weather on Thursday* |
| **Prediction** | ähnliches wetter auch am donnerstag | *similar weather on Thursday* |
| **Reference** | und nun die wettervorhersage für morgen montag den achtzehnten april | *and now the weather forecast for tomorrow, Monday, April 18th* |
| **Prediction** | und nun die wettervorhersage für morgen montag den achtzehnten april | *and now the weather forecast for tomorrow, Monday, April 18th* |
| **Reference** | am tag sechzehn grad an der ostsee und bis zu siebenundzwanzig grad am oberrhein | *during the day sixteen degrees on the Baltic Sea and up to twenty-seven degrees on the Upper Rhine* |
| **Prediction** | am tag sechzehn grad an der ostsee und bis zu siebenundzwanzig grad am oberrhein | *during the day sixteen degrees on the Baltic Sea and up to twenty-seven degrees on the Upper Rhine* |
| **Partial-Success Cases** | | |
| **Reference** | im süden entladen sich später hier und da heftige gewitter | *In the south, heavy thunderstorms will occur here and there later* |
| **Prediction** | im südosten sind noch kräftige gewitter möglich | *Strong thunderstorms are still possible in the southeast* |
| **Reference** | gleichzeitig bleibt es unter hochdruckeinfluss weitgehend trocken und teilweise wird es recht freundlich morgen früh | *Otherwise, under the influence of high pressure, it will remain mostly dry and partly pleasant during the day* |
| **Prediction** | ansonsten bleibt es unter hochdruckeinfluss meist trocken und tagsüber auch teilweise freundlich | *At the same time, under the influence of high pressure, it will remain largely dry and partly it will be quite pleasant tomorrow morning* |
| **Reference** | am mittwoch breiten sich von nordwesten neue wolken mit regen oder schauern aus örtlich sind auch gewitter dabei | *On Wednesday, new clouds with rain or showers will spread from the northwest, with thunderstorms in some places.* |
| **Prediction** | am mittwoch ziehen von nordwesten erneut teilweise kräftige schauer und gewitter heran die gebietsweise unwetterartig sein können | *On Wednesday, partly heavy showers and thunderstorms will again approach from the northwest, which may be severe in some areas* |
| **Failure Cases** | | |
| **Reference** | mit der leicht kühleren luft dann bis vierundzwanzig grad in der nordhälfte | *with the slightly cooler air then up to twenty-four degrees in the northern half* |
| **Prediction** | in den nächsten tagen bis vierundzwanzig grad an der ostseeküste | *in the next few days up to twenty-four degrees on the Baltic Sea coast* |
| **Reference** | die temperaturen sinken auf werte zwischen dreizehn und fünf grad | *temperatures drop to between thirteen and five degrees* |
| **Prediction** | und die temperaturen die klettern weiter auf höchstwerte zwischen dreizehn und fünfundzwanzig grad | *and the temperatures continue to climb to highs between thirteen and twenty-five degrees* |

Table 3. **Qualitative Examples of Translation Performance.** We show representative outputs from our model across three categories: *success* (exact match with the reference), *partial success* (semantically similar but not identical), and *failure* (significant deviation from the reference). English translations are also provided in italics at the right column for convenience.

## References


[1] Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. Improving 3d pose estimation for sign language. In *ICASSP*, pages 1–5. IEEE, 2023. 1

[2] Jianhe Low, Harry Walsh, Ozge Mercanoglu Sincan, and Richard Bowden. Hands-on: Segmenting individual signs from continuous sequences. In *FG*, 2025. 1

[3] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *CVPR*, pages 9826–9836, 2024. 1