Appendix

We begin by providing a detailed analysis of the sign language datasets used in this work in Appendix A. In Appendix B, we discuss additional implementation details, including keypoint preprocessing and the network architecture. Further qualitative results are presented in Appendix C. Lastly, we address the potential negative societal impacts of our work in Appendix D.

A. Statistics of Sign Language Datasets

We provide details on two sign language datasets used in this work: PHOENIX14T and How2Sign, as summarized in Tab. 7. PHOENIX14T is a German Sign Language (DGS) dataset focused on the specific domain of weather forecasting. It features a relatively small vocabulary of 3K words and concise video clips averaging 116 frames in length. The dataset includes 7,096 training samples, 519 validation samples, and 642 test samples, all with gloss annotations. Because it is tailored for domain-specific tasks, PHOENIX14T offers clear and repetitive patterns, making it ideal for translation and recognition tasks within weather-related contexts. On the other hand, How2Sign is an American Sign Language (ASL) dataset within the instructional domain. This dataset is significantly larger and more diverse, with a vocabulary of 16K words and an average video length of 173 frames. It contains 31,128 training samples, 1,741 validation samples, and 2,322 test samples, though it lacks gloss annotations. The broader and more complex nature of How2Sign makes it well-suited for general sign language processing tasks, particularly those that require an understanding of diverse and intricate sign sequences.

B. More Implementation Details

B.1. Keypoint Preprocessing

In preprocessing the keypoints, we first center and normalize them based on the shoulder joint, ensuring that the length of the shoulder is normalized to 1, following [77]. However, off-the-shelf extraction models, such as OpenPose [9], do not always yield consistently high-quality keypoint data. To address this issue, we implemented an additional step to filter out noisy frames—specifically, those with missing or misplaced joints—to ensure data quality and consistency [44, 73]. A frame is considered noisy if the joint distance between consecutive frames exceeds a certain threshold. To detect such frames, we first calculate the differences between consecutive frames as:

$$X_{\text{diff}} = \{x_t - x_{t-1}\}_{t=1}^{T}, \quad x \in \mathbb{R}^{V \times C} \tag{7}$$

for $t = 1, ..., T$, where T is the total number of frames, V is the number of vertices (joints), and C represents the coordinates (x and y). We then compute the Euclidean distance

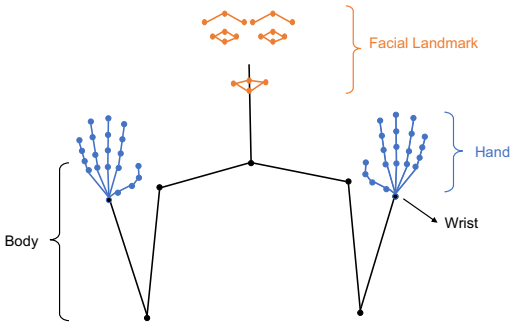


Figure 6. A visual abstract of the keypoints used: 73 keypoints in total, including 19 for the face, 23 for each hand, and 8 for the upper body.

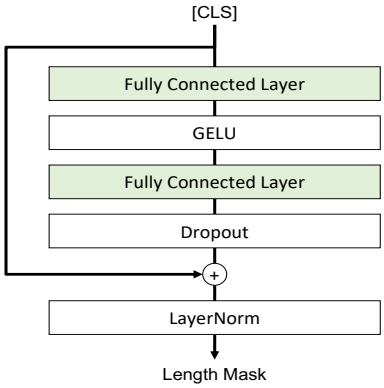


Figure 7. An overview of the length regulator.

on the difference X_{diff} as:

$$\|X_{\text{diff}}(t, v)\| = \sqrt{\sum_{c=1}^{C} (X_{\text{diff}}(t, v, c))^2}, \tag{8}$$

where $X_{\text{diff}}(t, v, c)$ denotes the difference at time t for vertex v and coordinate c. Next, we calculate the mean Euclidean distance across all joints between consecutive frames:

$$\overline{D}(t) = \frac{1}{V} \sum_{v=1}^{V} \|X_{\text{diff}}(t, v)\| . \tag{9}$$

Frames where $\overline{D}(t)$ exceeds a predefined threshold, empirically set to 400, are considered noisy and are subsequently removed.

B.2. Network Details

We employ a Transformer encoder and decoder architecture [64] for our models, largely following the configuration of Joint-SLT [8]. Our framework is optimized in two stages: (i) pretraining SignMAE and (ii) task-specific mapping for SLT and SLP. During pretraining, we empirically set a 25% random mask to extract spatio-temporal features. Typically,

Dataset	Lang	Vocab	Train / Valid / Test	Avg. No.F	Gloss	Domain
PHOENIX14T [7]	DGS	3K	7,096 / 519 / 642	116	✓	Weather Forecast
How2Sign [21]	ASL	16K	31,128 / 1,741 / 2,322	173	✗	Instructional

Table 7. Statistics of Sign Language Datasets. Avg. No.F means average number of frames.

MAE mask a large portion of input data (e.g., 75%) to enhance the robustness of learned representations. However, in our case, where both the encoder and decoder are used for SLT and SLP tasks, the 25% random mask provides a balanced approach. It offers a sufficient level of masking while retaining enough visible context for effective reconstruction.

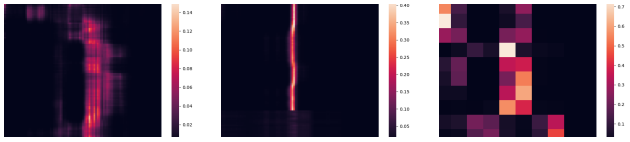
To encode spoken language sentences, we employed DistilBERT[2]. The length regulator module, illustrated in Fig. 7, comprises two fully connected linear layers with GELU activation [28], Dropout [58], and Layer Normalization [2]. We used a learnable query size of 128 as input to the decoder for generating gloss-level representations. A batch size of 64 was maintained across all tasks, including both pretraining and task-specific mapping. The model was optimized using the AdamW optimizer [46] with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a weight decay of 1e-3. The learning rate schedule followed a cosine decay, with a peak learning rate of 1e-4 and a linear warmup over 10K steps, decaying to a minimum learning rate of 5e-5.

We trained SignMAE for 50 epochs on a single NVIDIA A100 GPU, completing the process in under 24 hours. For sign-to-text mapping, the model was trained for 75 epochs, with early stopping applied when no further improvement in BLEU-4 was observed. Training took approximately 1 hour for PHOENIX14T and 3 hours for How2Sign. We used a beam size of 5. For text-to-sign mapping, training was completed in 100 epochs, with early stopping based on back-translated BLEU-4. The process was completed in 1 hour for PHOENIX14T and 7 hours for How2Sign.

C. More Results

C.1. Visualization of the Attention Map

Building on the findings of [75], which demonstrated that glosses provide alignment information, we present visualizations of the attention maps from three different SLT methods: Joint-SLT with and without the glosses, and our method. As shown in Fig. 8a, the use of the glosses clearly helps the model focus on more important local areas by providing essential alignment information. By contrast, as shown in Fig. 8b, removing the gloss supervision signal causes the model to struggle in identifying the correct regions. However, as shown in Fig. 8c, our method maintains attention on key regions effectively, performing similarly to the model using the glosses, thereby showing its potential as a gloss replacement.



(a) W/ Gloss (b) W/o Gloss (c) Ours

Figure 8. Visualization of attention maps in the shallow encoder layer of Joint-SLT with and without glosses, alongside our method. The comparison highlights how each model focuses on different aspects of the input data.

C.2. Translation Results

We provide additional translation examples in Tab. 8. The results demonstrate that our method consistently delivers accurate and semantically correct translations, while other baselines struggle to capture the correct semantic meaning.

C.3. Production Results

We provide additional production examples in Figs. 9 and 10. The results demonstrate that our method consistently delivers accurate and semantically correct signs, while other baselines struggle to produce correct signs.

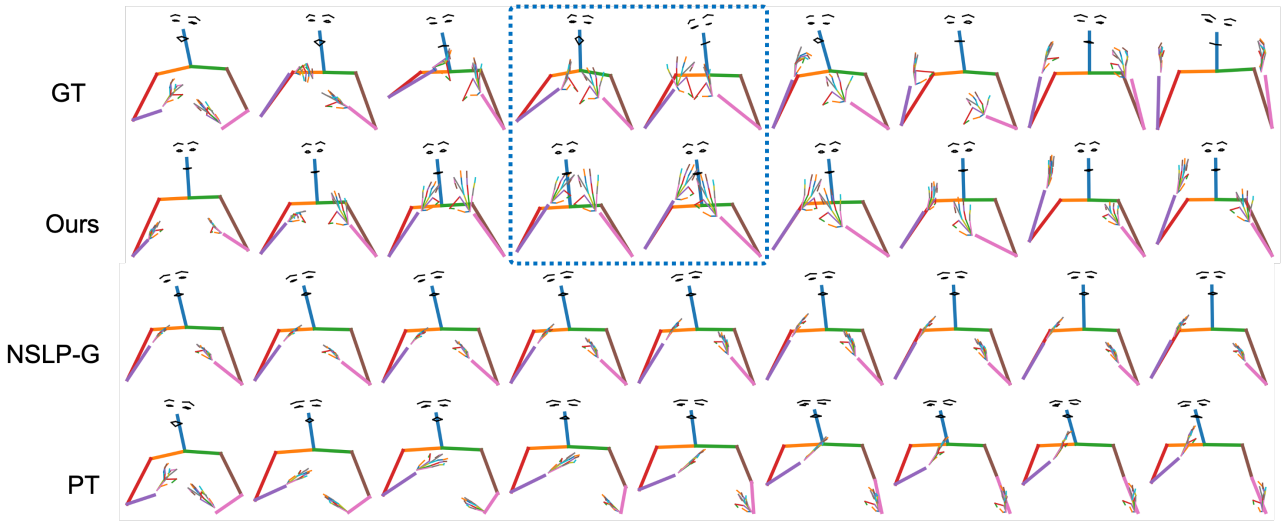
D. Potential Negative Societal Impact

On the positive side, the development and refinement of frameworks such as UniGloR, which aim to replace traditional glosses, offer transformative potential, particularly for advancing machine learning applications for Deaf and Hard-of-Hearing communities. By applying this framework to bidirectional sign language translation systems, we can bridge the communication gap between the hearing and Deaf communities. This breakthrough holds the promise of creating more inclusive educational, professional, and social environments, reducing the marginalization of Deaf and Hard-of-Hearing individuals. However, a potential negative impact arises from the datasets we use. Publicly available datasets, such as PHOENIX14T [7] and How2Sign [21], may contain identifiable information, raising concerns about personal privacy. To mitigate these concerns, we work with extracted keypoints from sign videos, ensuring that no personally identifiable information is included in our training process.

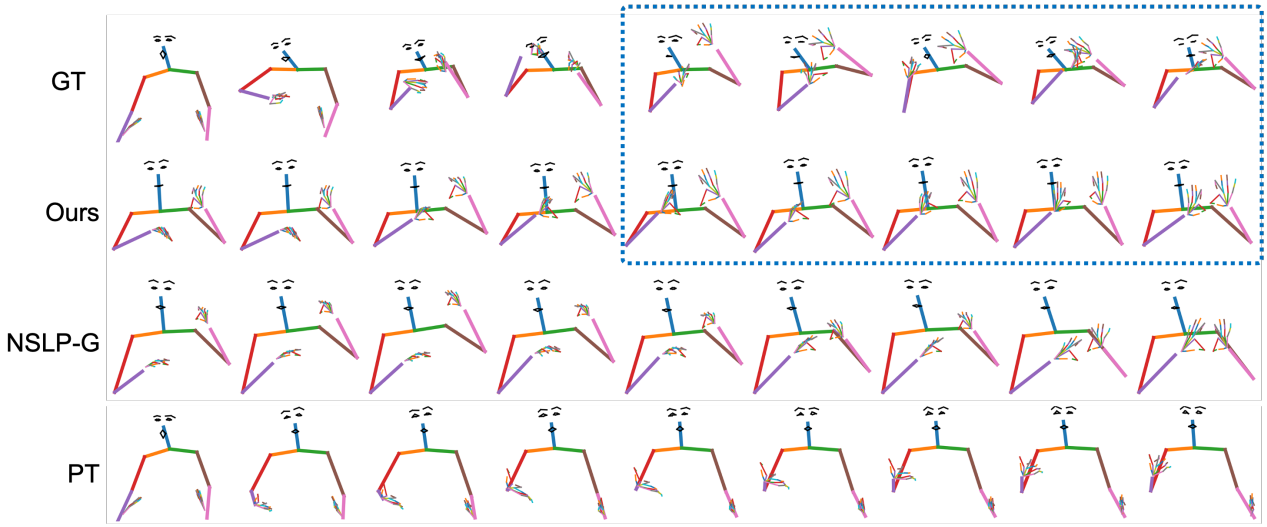
[2] https://huggingface.co/M-CLIP/M-BERT-Distil-40

	PHOENIX14T
Ground Truth:	und nun die wettervorhersage für morgen montag den achtundzwanzigsten november. (*And now the weather forecast for tomorrow monday the twenty-eighth of november.*)
Joint-SLT [8]	und nun die wettervorhersage für morgen donnerstag den sechsundzwanzigsten juli. (*And now the weather forecast for tomorrow thursday twenty-sixth of July.*)
ConSLT [22]	und nun die wettervorhersage für morgen samstag den sechsundzwanzigsten juli. (*And now the weather forecast for tomorrow saturday the twenty-sixth of July.*)
Ours:	und nun die wettervorhersage für morgen donnerstag den siebenundzwanzigsten november. (*And now the weather forecast for tomorrow thursday the twenty-seventh of November.*)
Ground Truth:	am samstag ist es in der südhälfte freundlich in der nordhälfte einzelne schauer. (*On Saturday, it will be friendly in the southern half and a few showers in the northern half.*)
Joint-SLT [8]	am samstag ist es im nordwesten freundlich (*On Saturday it is friendly in the northwest.*)
ConSLT [22]	auch am samstag ist es im südosten freundlich. (*Even on Saturday it is friendly in the southeast.*)
Ours:	am samstag scheint im süden oft die sonne und wolken sonst einzelne schauer. (*On Saturday, the sun often shines in the south and clouds otherwise a few showers.*)
Ground Truth:	am sonntag ziehen von nordwesten wieder schauer und gewitter heran. (*On Sunday, showers and thunderstorms will move in again from the northwest.*)
Joint-SLT [8]	am samstag neun grad an der küste. (*On Saturday, nine degrees on the coast.*)
ConSLT [22]	am sonntag ziehen dann von westen wieder neue niederschläge heran. (*On Sunday, new precipitation will move in from the west.*)
Ours:	am sonntag zieht von nordwesten wieder etwas regen heran. (*On Sunday, some rain will move in again from the northwest.*)
	How2Sign
Ground Truth:	Today, I'm going to show you how to play three card poker.
Joint-SLT [8]	So what we're going to do is we're going to take our sponge.
ConSLT [22]	You want to make sure that you have your seatbelt buckled.
Ours:	Today I'm going to show you how to do this.
Ground Truth:	Now we're going to talk specifically about the medication that is going to be used with this machine.
Joint-SLT [8]	So what we're going to do is we're going to start on the bottom here and we're going to do it in slow-motion.
ConSLT [22]	So what we're going to do is we're going to start by pruning on the bottom and we're going to go over the top.
Ours:	In this clip, we're going to talk about the proper way to get rid of the skin.

Table 8. Comparison of the translation results compared to baselines. Correctly translated 1-grams are highlighted in blue, and semantically correct translations are highlighted in green.

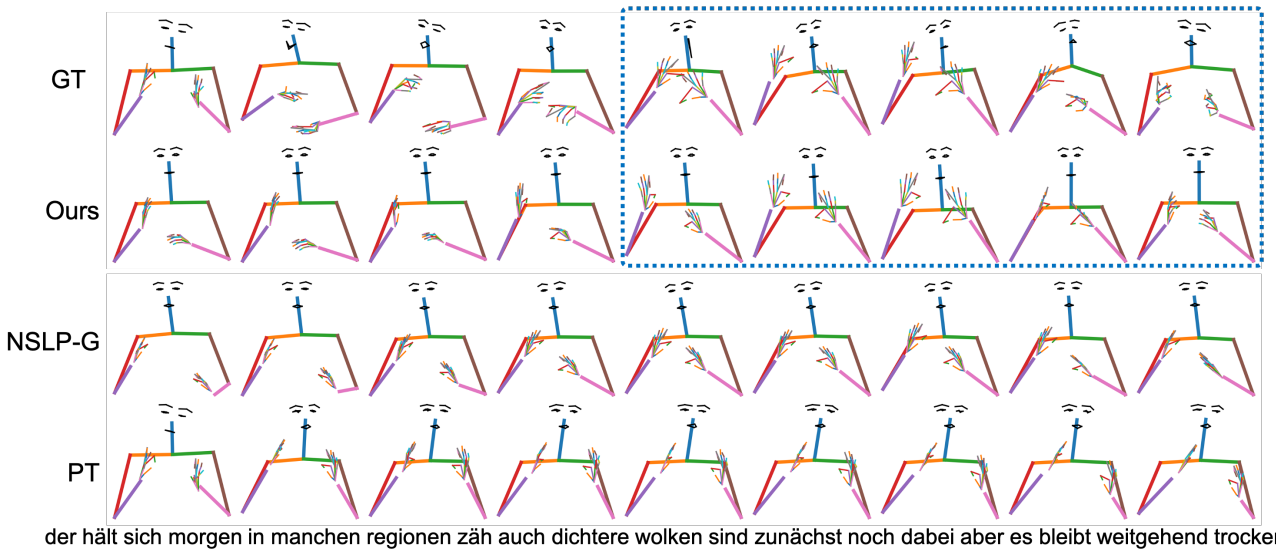


das kräftige mittelmeertief dehnt seinen einfluss weiter nach norden aus und sorgt morgen
in fast ganz deutschland für einen wolkenverhangenen Himmel.
(*The strong Mediterranean low extends its influence further north and will ensure cloudy skies in almost all of Germany tomorrow.*)

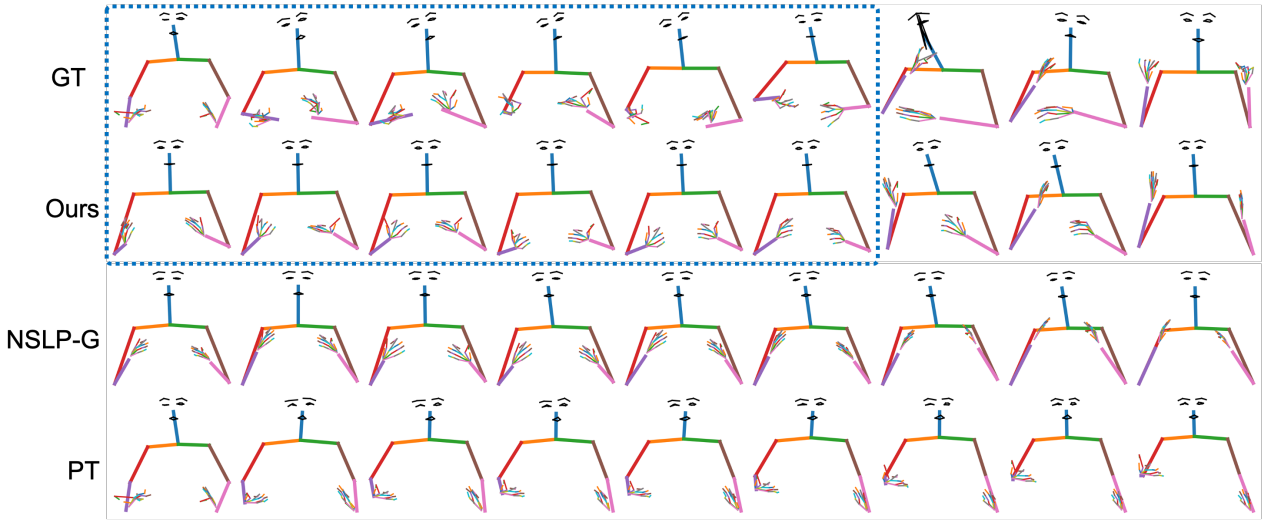


die wolken eines tiefs bei schottland erreichen uns morgen von westen her.
(*The clouds of a deep near Scotland reach us from the west tomorrow.*)

Figure 9. Additional visualization examples generated by our method, NSLP-G, and PT. Frames were uniformly selected, with dashed boxes highlighting areas where our method produced more accurate signs.



der hält sich morgen in manchen regionen zäh auch dichtere wolken sind zunächst noch dabei aber es bleibt weitgehend trocken.
(Tomorrow he stays tough in some regions, even denser clouds are still there, but it remains largely dry.)



der wind weht schwach bis mäßig am meer auch frisch.
(*The wind also blows weak to moderately on the sea.*)

Figure 10. Additional visualization examples.