SLRTP2025 Sign Language Production Challenge: Methodology, Results, and Future Work

Harry Walsh[1*], Ed Fish[1*], Ozge Mercanoglu Sincan[1*], Mohamed Ilyes Lakhal[1*], Richard Bowden[1*],
Neil Fox[2], Bencie Woll[3], Kepeng Wu[4], Zecheng Li[4], Weichao Zhao[4], Haodong Wang[4],
Wengang Zhou[4], Houqiang Li[4], Shengeng Tang[5], Jiayi He[5], Xu Wang[5], Ruobei Zhang[5],
Yaxiong Wang[5], Lechao Cheng[5], Meryem Tasyurek[6], Tugce Kiziltepe[6], Hacer Yalim Keles[6]
[1]University of Surrey, [2]University of Birmingham, [3]University College London,
[4]University of Science and Technology of China, [5]Hefei University of Technology, [6]Hacettepe University
{harry.walsh, edward.fish, o.mercanoglusincan, m.lakhal, r.bowden}@surrey.ac.uk
[*] denotes key workshop and challenge organisers.

Abstract

*Sign Language Production (SLP) is the task of generating sign language video from spoken language inputs. The field has seen a range of innovations over the last few years, with the introduction of deep learning-based approaches providing significant improvements in the realism and naturalness of generated outputs. However, the lack of standardized evaluation metrics for SLP approaches hampers meaningful comparisons across different systems. To address this, we introduce the first Sign Language Production Challenge, held as part of the third SLRTP Workshop at CVPR 2025. The competition's aims are to evaluate architectures that translate from spoken language sentences to a sequence of skeleton poses, known as Text-to-Pose (T2P) translation, over a range of metrics. For our evaluation data, we use the RWTH-PHOENIX-Weather-2014***T*** *dataset, a German Sign Language - Deutsche Gebärdensprache (DGS) weather broadcast dataset. In addition, we curate a custom hidden test set from a similar domain of discourse.*

This paper presents the challenge design and the winning methodologies. The challenge attracted 33 participants who submitted 231 solutions, with the top-performing team achieving BLEU-1 scores of 31.40 and DTW-MJE of 0.0574. The winning approach utilized a retrieval-based framework and a pre-trained language model. As part of the workshop, we release a standardized evaluation network, including high-quality skeleton extraction-based keypoints establishing a consistent baseline for the SLP field, which will enable future researchers to compare their work against a broader range of methods.

1. Introduction

Sign languages, like spoken languages, are complex systems with distinct grammar and vocabularies. They are independent and fully fledged languages with a unique structure, using manual and non-manual features asynchronously to convey information [41]. Manual features can be defined as the physical motion, location, and shape of the hands and arms, while non-manual features include facial expressions, head movements, and body posture. Translating between spoken and signed languages presents a significant communication challenge, typically requiring expert human interpreters rather than simple word-to-sign substitution. However, the scarcity of interpreters limits access to information for the Deaf community. Sign Language Production (SLP), the task of generating sign language from spoken language inputs, has the potential to be part of a solution to improving accessibility.

Over the last three decades, research into SLP has made significant progress [42]. While early approaches utilized graphical avatars and rule-based translation systems [2, 10–12, 55], these architectures often produced robotic and unnatural movement. However, more recent progress in deep learning-based methods has significantly enhanced the realism and naturalness of sign language generation [7, 33, 35, 47, 49, 50]. Approaches to SLP commonly decompose the task into several stages, using intermediate representations such as linguistic annotation [39, 46, 48], skeleton pose [33, 49], and parametric human models [1, 40, 54].

Despite these advancements, progress in the field has been impeded by the absence of standardized evaluation metrics, which hinders meaningful comparisons across SLP systems. To address this issue, and as part of the SLRTP

Workshop[1], a sign language production challenge was held, attracting 33 participants who submitted 231 solutions[2]. In this manuscript, we present the approaches from the top three performing teams. The competition focused on developing robust systems for spoken language to sign language translation. As part of the evaluation pipeline, we propose a novel metric, total distance, to help score the expressiveness of the productions. Given previous work has noted issues with regression to the mean, the proposed metric provides an improved qualitative measure for translation quality. We release the evaluation pipeline[3], with the hope that this helps establish a consistent baseline for SLP work, enabling future researchers to conduct comprehensive comparisons with a wider range of methods.

The rest of this paper is organized as follows. Section 2 reviews related work in the field of Sign Language Production (SLP). Sections 3 and 4 describe the design and dataset used in the challenge. Section 5 details the evaluation protocol. Section 6 outlines the methodology used by the top three teams, followed by Section 7, which presents the results. Finally, Section 8 concludes the paper and suggests directions for future work.

2. Related Work

2.1. Sign Language Production

The shortage of qualified sign language interpreters has motivated the development of SLP systems. Computational sign language research started in the early 90s [42], with the first approaches being avatar-based [2, 10–12, 55]. However, the use of poor quality avatars has consistently received criticism from the Deaf community due to their unrealistic appearance and robotic motion, which limits comprehension [15]. Some of these systems looked for "legal" phrases in the input text and then mapped them to pre-defined sign language animations [10]. Another approach uses Motion Capture (MoCap) to incorporate more fluid motion dynamics, but this requires specialised capture equipment which then limits the vocabulary size [14].

The field then progressed to Statistical Machine Translation (SMT), allowing for automatic rule learning and improved generalizability [3, 22, 27]. These models were able to use the context of the input text to solve for ambiguity. However, such approaches require handcrafted features to be extracted before being used in an ensemble of models.

The introduction of deep learning has resulted in more data-driven approaches, that can learn the mapping between spoken and sign languages. The first deep SLP pipeline broke down the task into three steps: first, a translation from Text-to-Gloss (T2G), followed by a Gloss-to-Pose (G2P) look-up, and finally a Pose-to-Sign (P2S) video generation step [38]. However, this approach was unable to blend the motion between signs, and due to the low-resolution output key features such as facial expressions were lost.

Later approaches attempted to synthesize poses from textual input. Zelinka et al. [53] utilized a Recurrent Neural Network (RNN) to predict a 7-frame pose sequence for each word in the sentence. As a result, the length of each prediction is dependent on the number of words in the sentence. This limited the realism of the outputs. After, Saunders et al. [33] introduced the progressive transformer, the first to learn a direct mapping between text and pose using a transformer architecture. This approach was able to generate more realistic outputs and deal with the variable length nature of sign language. Further improvements were achieved by adding adversarial training and a Mixture Density Network (MDN) [31, 34]. A non-autoregressive transformer was also introduced to improve the speed of the model [19]. Recently, diffusion has been applied to the tasks [7, 44] and has shown improvements, but this work relies on linguistic annotation to guide the process. Similar to most methods discussed so far, large performance increases can be gained by introducing gloss[4] annotations [19, 31, 33, 34]. However, using gloss is a major limiting factor when scaling to larger domains of discourse.

Models that attempted to directly regress a pose sequence from text input often struggled with regression to the mean. This leads to the generation of less expressive outputs. To address this, some works have applied Vector Quantisation (VQ) to the task, where the model is used to learn a set of discrete codes that map to a small sequence of poses. This then serves as the lexicon for translation. Some of which predict the codes from text [20, 47], while others start from glosses [50].

Instead of learning units, others have used a dictionary of pre-recorded signs. They are guaranteed to be expressive and therefore produce comprehensible signing. Simply concatenating the signs can create an unnatural sequence, so Walsh et al. [49] used a 7-step pipeline to create smooth transitions, while others employed a transformer [54] or a diffusion model [43] for the task. Alternatively, Saunders et al. [36] proposed a novel keyframe selection network to learn the co-articulate between signs.

While the works discussed so far have employed skeleton poses as a representation for sign language the source can vary. Some researchers utilize MediaPipe [26], whereas others opt for OpenPose [6]. Significant variation exists across methods, with different authors using a different subset of keypoints. This has hindered meaningful comparison between these approaches.

[1] Workshop website: https://slrtpworkshop.github.io/

[2] Challenge website: https://www.codabench.org/competitions/4854/

[3] Challenge Evaluation Code: https://github.com/walsharry/SLRTP-Sign-Production-Evaluation/tree/main

[4] Gloss is the written words associated with a sign.

Beyond skeleton pose, human parametric models have also been used for the SLP task [1, 40, 54]. All of which leverage SMPL-X [25], a human model with parameters to describe the face, hands, and body. Other methods have directly predicted RGB video frames [16, 51]

A substantial body of research exists; however, direct comparison between these works is challenging due to the heterogeneity of output representations. We observe significant variability, particularly in skeleton pose representations. Furthermore, extraction and normalization techniques diverge considerably across studies. This inconsistency impedes the ability to discern whether performance improvements stem from novel algorithmic contributions or advancements in skeleton extraction and normalization methodologies. This issue is compounded by the limited public availability of codebases, which renders direct comparison between state-of-the-art approaches exceedingly difficult and highlights the need for more standardized evaluation metrics for SLP.

3. Challenge Design

The objective of this challenge was to generate continuous sign language sequences from spoken language inputs. We utilized a skeleton pose representation for sign language. As seen in the literature, this serves as a common intermediate representation and has been shown to be capable of driving photo-realistic signer generation [13, 29, 32].

The challenge comprised two phases. Initially, during the development phase, teams were provided with train and dev splits from the RWTH-PHOENIX-Weather-2014**T** (PHOENIX14**T**) dataset [4]. Throughout this phase, participants could submit solutions to the test set via the online platform, receiving evaluation scores for their skeleton pose predictions. Subsequently, in the final week, the competition transitioned to the test phase. Participants were presented with 500 spoken language sentences from the hidden test set and tasked with submitting their model's predictions. The competition was held on the Codabench platform [52], an open-source framework for running competitions, enabling the automated evaluation of code and results. The competition spanned 49 days, starting on January 13, and finishing on March 3, 2025. During this time, 33 participants contributed 231 solutions. During the development phase, participants were limited to 100 submissions per day, or a total of 3,000 submissions throughout the competition. During the final test phase, participants were limited to just three submissions. This constraint was implemented to minimize the likelihood of participants overfitting to specific data portions or employing random initializations to achieve performance gains.

In the final stage, six teams submitted solutions that outperformed the baseline approach. The top-performing teams were requested to submit a fact sheet detailing their approach. The teams with the highest scores, and who shared information about their code and methodology, were selected as the winners.

For the final ranking, in Sec. 7, we employ Pareto dominance. A solution dominates another if it performs at least as well in all metrics and demonstrably better in at least one. This method ensures that no single metric is arbitrarily favoured, providing a balanced and fair evaluation.

4. Challenge Datasets

For the SLRTP 2025 CVPR challenge, we use the RWTH-PHOENIX-Weather-2014T (PHOENIX14**T**) dataset [4] and a custom hidden test set. Here we provide details on each.

4.1. PHOENIX14T

The PHOENIX14**T** dataset is one of the most widely used benchmarks for research in sign language recognition and translation. It contains continuous sign language videos along with their corresponding gloss annotations and spoken language subtitles. The dataset is derived from weather forecast broadcasts in German Sign Language (DGS).

The dataset is split into training, development, and test sets. The training set contains 7096 videos, the development set contains 519 videos, and the test set contains 642 videos.

4.2. Hidden Test Set

The hidden test set for this challenge was sourced from the Phoenix broadcast channel, consistent with the original PHOENIX14**T** dataset. This data was collected as part of the EASIER project[5]. However, unlike the original dataset, it contains minimal manual annotation. Therefore, we first searched the subtitles for sections most relevant to weather-related discussions, to maintain the original domain of discourse. We then cropped these sections from the original videos. To ensure that the selected segments were indeed from weather broadcasts, we manually verified a subset of the total videos. Finally, we randomly selected 500 sentences to form the hidden test set.

4.3. Skeleton Representation

For each video, we extract Mediapipe holistic keypoints and use the approach from Ivashechkin *et al.* [21], to uplift the predictions to 3D. This optimization leverages a neural network, informed by human body physical constraints, to predict 3D skeleton joint angles from 2D keypoints. These angles are then used to apply a canonical skeleton, thereby ensuring consistent bone lengths across all signers. This provides 178 keypoint representation: 21 keypoints for each hand, 128 keypoints for the face, and, 8 keypoints for the

[5] https://www.project-easier.eu/

body. The face is a subset of the 468 keypoint representations from Mediapipe, we do this for computational efficiency. We then normalise the skeleton such that the neck is at the origin and the body as is fixed on the xy plane. An example of the skeleton extraction can be seen in Fig. 1.

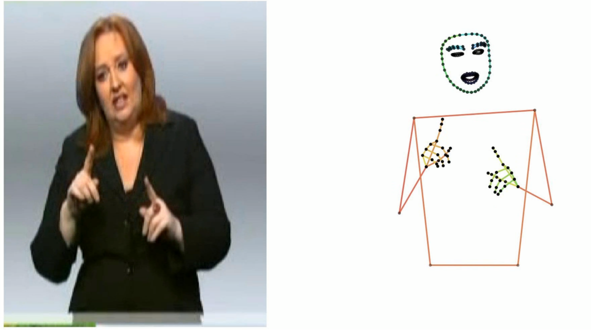

Figure 1. Skeleton extraction example, (left: original signer, right: 178 keypoint skeleton.

5. Evaluation Metrics

Here we discuss the evaluation protocol used for the SLP challenge, which is also publicly available[6]. The evaluation is comprised of text-based and pose-based metrics. Text-based metrics employ a back-translation model to convert the skeletal representation back into spoken language. Specifically, a "Sign Language Transformer" [5] is utilized. The input layer is adapted to match the dimensionality of the keypoints. Both the encoder and the decoder component comprise three layers, 8 attention heads, with an embedding and feedforward dimensions of 256 and 512, respectively. Whereas the pose-based metric is calculated using the skeleton itself.

5.1. Text-based

BLEU: The Bilingual Evaluation Understudy (BLEU) score [28] compares a given sentence to that of a set of reference sentences by calculating the precision of word-level n-grams. Here we use n-grams one to four.

CHRF: The CHaRacter-level F-score (CHRF) [30] calculates the F-score based on the precision and recall of character-level n-grams between two sentences.

ROUGE: The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score [24] measures the overlap of n-grams, word sequences, and word pairs between two sentences.

WER: The Word Error Rate (WER) calculates the number of errors (insertions, deletions, and substitutions) divided by the total number of words in the reference text.

5.2. Pose-based

DTW MJE: The Dynamic Time Warping Mean Joint Error (DTW MJE) is a metric used to evaluate the similarity between two sequences of skeleton poses. It calculates the average error between corresponding joints in the predicted and ground truth poses, considering temporal alignment.

Total Distance: This metric measures the overall distance the signer's hands have moved in 3D space. This is to judge how expressive the productions are. The prediction is normalized by the ground truth distance, therefore, a score of 1 is optimal.

5.3. Baseline Model

For a baseline method, we use the publicly available "Progressive transformer" by Saunders *et al.* [33]. The model can translate from text to continuous 3D sign pose sequences in an end-to-end manner. The model is based on the transformer architecture, which has been shown to be effective for sequence-to-sequence tasks such as machine translation. By introducing a novel counter decoding technique, the model can generate continuous pose sequences of variable lengths. We train a model for 300 epochs with an Adam optimizer with a learning rate of 0.001. In line with the original paper, both the encoder and decoder contain 2 layers and 8 heads with an embedding size of 512.

6. Methodology

In this section, we introduce the three top-performing approaches in the SLRTP 2025 Sign Language Production Challenge. *Team 1* employs a retrieval-based pipeline centered on gloss annotations, *Team 2* leverages a generative diffusion-based model, and *Team 3* presents a gloss-free transformer architecture with an autoencoder for latent pose embeddings. Tab. 1 summarizes the key characteristics of each method.

Participant	Team-1: USTC-MoE	Team-2: hfut-lmc	Team-3: Hacettepe
Model architecture	Rule based	Diffusion	Transformer
Pre-trained models	✓	-	✓
No. Trainable Parameters	355.9M	125.8M	26M
External Datasets	-	-	-
Gloss Information	✓	-	-
Data Augmentation	-	-	-
Handcrafted Features	-	-	-
Motion Constraints	-	✓	-
Optimization Metric	BLEU and DTW	DTW	BLEU-4

Table 1. General information about the top-3 winning approaches.

The following subsections detail each team's respective methodologies, including training setups and key design choices.

[6] Evaluation code: https://github.com/walsharry/SLRTP-Sign-Production-Evaluation

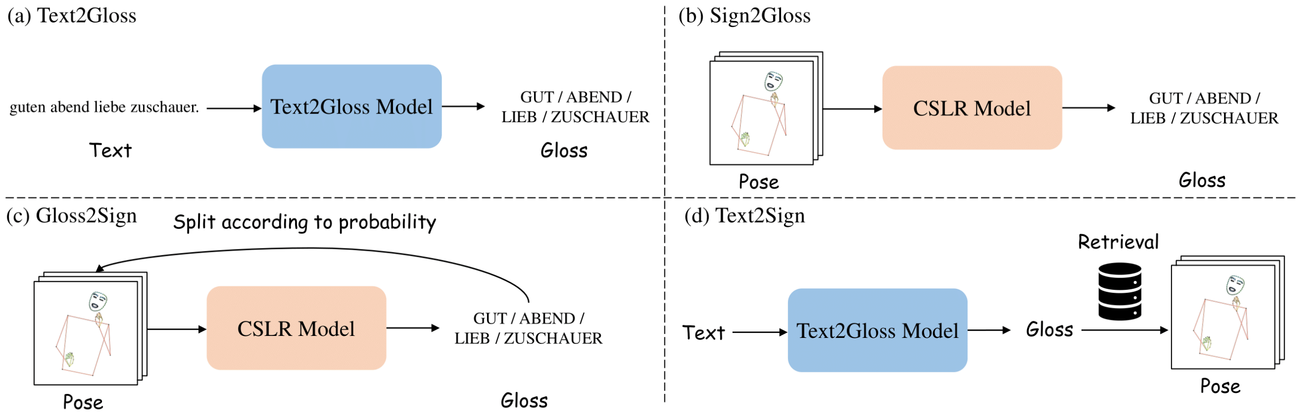

Figure 2. Overview figure of the 1st place method (USTC-MoE).

6.1. Team 1: USTC-MoE (1st Place)

Overall Pipeline. *Fig. 2* (adapted from the team's fact sheet) outlines the top-1 approach. It is a retrieval-based framework that connects spoken language input to a large dictionary of 3D pose segments. The method consists of four key modules: *Text2Gloss*, *Sign2Gloss*, *Gloss-Pose Dictionary Construction*, and *Sign Retrieval*.

(a) Text2Gloss. A multilingual pretrained language model, specifically XLM-R [9], is employed to convert each spoken language sentence into a gloss sequence. This model is fine-tuned with a cosine annealing schedule for 40 epochs using Adam (weight decay 10^{-3}, learning rate 10^{-5}).

(b) Sign2Gloss. They then train a Continuous Sign Language Recognition (CSLR) model (based on TwoStream-SLT [8]), which outputs a gloss label for every frame of the 3D pose sequence. This step provides a robust mapping from skeleton pose to discrete glosses.

(c) Gloss-Pose Dictionary Construction. After the CSLR network is trained, it is used to label and segment all training poses. Each gloss is therefore associated with its corresponding 3D sub-pose sequence, producing a large *Gloss-Pose* dictionary of short signing segments.

(d) Text2Sign. At inference, the model (i) translates text to gloss, (ii) retrieves each sub-pose from the dictionary based on the gloss, and (iii) concatenates these short pose segments to form the final sign pose sequence. They report a total parameter count of about 355.9 M (dominated by the pretrained text model). No additional data augmentation is performed. The reliance on real pose segments ensures high fidelity and natural transitions across the retrieved signs.

Discussion. By grounding each gloss in an actual segment of human motion, this retrieval-based approach bypasses the complexities of motion synthesis and guarantees plausible, high-quality poses for individual signs. Concatenating these sub-pose sequences requires robust gloss alignment, achieved via the CSLR model. Its simplicity and reliability, coupled with strong text-to-gloss translation, led to first-place performance.

6.2. Team 2: hfut-lmc (2nd Place)

Overall Pipeline. Unlike the retrieval-based strategy from Team 1, hfut-lmc [17] proposes a fully generative diffusion-based framework, illustrated in *Fig. 3*. Referred to as *Text-Driven Conditional Diffusion Model (TCDM)*, their method learns an end-to-end mapping from textual input to sign language pose sequences without leveraging gloss-level supervision or large retrieval dictionaries.

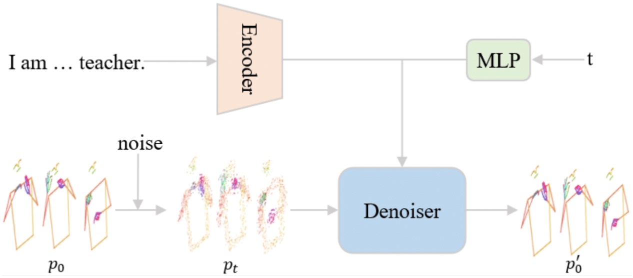

Figure 3. Overall framework of the 2nd place method (hfut-lmc).

Forward and Reverse Processes. Following the standard diffusion paradigm [18, 37], the forward process gradually adds Gaussian noise to the ground-truth 3D pose sequence p_0. After t steps, the noisy sequence p_t is:

$$p_t \;=\; \gamma_t \, p_0 \;+\; \sigma_t \, \epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ and $\gamma_t^2 + \sigma_t^2 = 1$. During training, the model learns a denoiser $\mathcal{D}\big(p_t,\, g\big)$ that removes noise from p_t, guided by a condition g derived from the text encoder

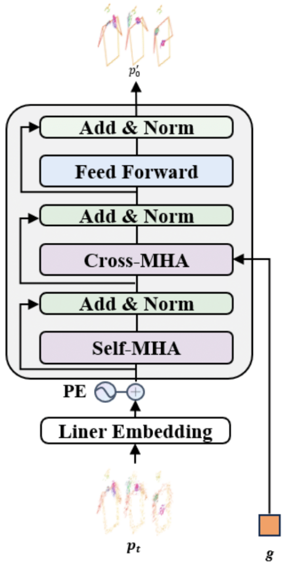

Figure 4. Detailed implementation details of the denoiser $\mathcal{D}$, from the 2^{nd} place method (hfut-lmc).

plus the timestep t:

$$p'_0 \;=\; \mathcal{D}\big(p_t,\, g\big). \tag{2}$$

At inference, an initially random pose p_T is iteratively denoised over a small number (5) of DDIM sampling steps, culminating in a coherent sign language motion.

Denoiser Architecture. To clarify how $\mathcal{D}$ operates, hfut-lmc splits the procedure into several sub-stages (Fig. 4):

1. *Linear Embedding Layer (LE):* The noisy pose p_t is projected into a higher-dimensional space:

$$p_u \;=\; W^p\, p_t \;+\; b^p, \tag{3}$$

   where p_u is the embedded representation of p_t.
2. *Positional Encoding (PE):* A predefined sinusoidal encoding is added to inject temporal information:

$$\hat{p}_u \;=\; p_u \;+\; PE(n), \tag{4}$$

   where n indexes each frame in the sequence.
3. *Multi-Head Attention and Cross-Attention:* The embedded pose $\hat{p}_u$ interacts with the conditioning g (derived from the text encoder and the current diffusion timestep) via cross-attention, yielding updated pose features that finally yield a refined pose estimate $\hat{p}'_0$.

Loss Functions. To ensure the generated poses are both accurate and realistic, hfut-lmc combines two complementary objectives:

- *Joint Position Loss* $\mathcal{L}_{\text{joint}}$,

$$\mathcal{L}_{\text{joint}} \;=\; \frac{1}{J}\sum_{j=1}^{J} \big\| p_j - p'_j \big\|, \tag{5}$$

  enforcing alignment between predicted and ground-truth joint coordinates.
- *Bone Orientation Loss* $\mathcal{L}_{\text{bone}}$,

$$\mathcal{L}_{\text{bone}} \;=\; \frac{1}{B}\sum_{b=1}^{B} \big\| q_b - q'_b \big\|^2, \tag{6}$$

  promoting realistic limb orientations by comparing direction vectors for each bone.

The total training objective is

$$\mathcal{L}_{\text{total}} \;=\; \mathcal{L}_{\text{joint}} \;+\; \lambda\,\mathcal{L}_{\text{bone}}, \quad \lambda = 0.1. \tag{7}$$

They employ the text encoder architecture that is the same as the baseline method [33], but with 4 layers, 8 heads, and a 1024-dimensional embedding size. The forward diffusion steps is set to 1000 and is combined with a 5-step DDIM inference schedule. Training uses Adam [23] and a learning rate of 10^{-3}.

Discussion. By omitting gloss supervision, hfut-lmc avoids the expense of additional annotations. Their bone-orientation and joint positioning losses, ensure the poses represent natural articulations, mitigating regression-to-mean artifacts.

6.3. Team 3: Hacettepe (3^{rd} Place)

Overall Pipeline. Hacettepe presents a *gloss-free* transformer-based method that learns a compact, disentangled latent representation of sign pose sequences via an autoencoder. Channel-aware regularization guides text-to-pose mapping without gloss supervision, as shown in *Fig. 5*.

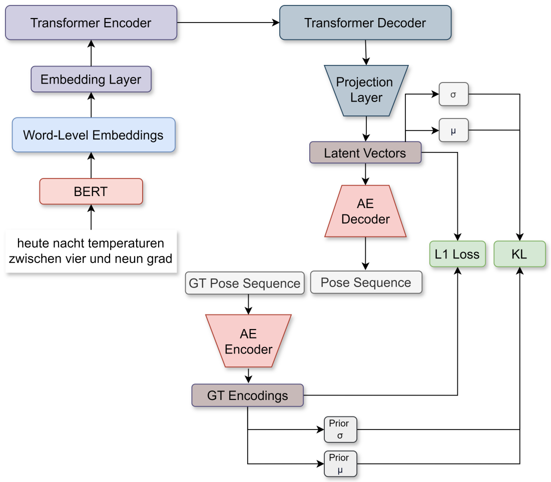

Figure 5. Workflow diagram for the 3^{rd} place method (Hacettepe).

Autoencoder for Latent Pose. A structurally disentangled pose autoencoder is first pre-trained to reconstruct 3D poses. Each 534-dimensional input is factorized into four

	RWTH-PHOENIX-Weather-2014**T** Test Set								
Teams	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CHRF	ROUGE	WER ↓	DTW-MJE ↓	Total Distance
Ground Truth	34.40	22.04	16.09	12.78	34.59	35.20	85.77	0.0000	1.000
Team 1 (USTC-MoE)	**34.85**	**21.96**	**15.65**	**12.06**	**36.83**	**36.59**	93.49	0.0448	1.631
Team 2 (hfut-lmc)	16.96	6.56	3.38	2.05	25.88	19.77	147.85	**0.0403**	2.512
Team 3 (Hacettepe)	30.44	17.75	12.42	9.59	29.70	30.64	**88.88**	0.0423	**0.798**
Progressive Transformer [33]	22.17	10.71	7.09	5.43	24.13	21.98	101.45	0.0418	0.257

Table 2. SLP Challenge Results on the RWTH-PHOENIX-Weather-2014**T** (PHOENIX14**T**) test set.

	Hidden Test Set								
Teams	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CHRF	ROUGE	WER ↓	DTW-MJE ↓	Total Distance
Ground Truth	37.94	19.87	10.67	5.90	30.64	38.60	101.25	0.0000	1.000
Team 1 (USTC-MoE)	**31.40**	**17.09**	**9.43**	**5.86**	**31.73**	**33.75**	109.38	0.0574	1.185
Team 2 (hfut-lmc)	30.54	16.22	9.33	5.66	30.17	32.92	107.93	**0.0492**	**0.971**
Team 3 (Hacettepe)	27.51	11.13	5.36	2.91	23.37	27.29	**105.49**	0.0531	0.761
Progressive Transformer [33]	18.33	4.99	1.74	0.78	21.65	21.10	141.93	0.0467	0.322

Table 3. SLP Challenge Results on the Hidden Test Set.

articulatory regions, face, body, left hand, and right hand, mapped to an 80-dimensional latent space. Region-specific L1 reconstruction and encoder-weight regularization ensure compact and semantically structured representations.

Transformer Architecture. A non-autoregressive transformer then predicts these latent embeddings from 768-dimensional sentence vectors obtained from a pretrained BERT model. Text embeddings are reduced to 512 dimensions before being processed by a 3-layer encoder and a 6-layer decoder to generate pose sequences in parallel, matching the autoencoder's latent space.

Reconstruction & Training. The transformer is trained with an L1 loss between predicted latent vectors and the autoencoder's ground-truth codes, ensuring realistic 3D motion reconstruction. In the second phase, channel-wise KL divergence promotes articulator-aware regularization. As it avoids gloss annotations, the model scales efficiently without costly linguistic labelling.

Discussion. Hacettepe's gloss-free, disentangled autoencoder-based approach offers a compact, interpretable representation for SLP. Unlike retrieval-based methods, it synthesizes entirely new sequences without gloss or a motion dictionary. Although bridging text to a learned pose space can present challenges for highly nuanced signs, the method demonstrates competitive accuracy and efficiency, securing third place in the challenge. Further details and extended results are provided in [45].

7. Challenge Results

On the hidden test set, all three teams significantly outperformed the baseline method. The top-performing team achieved an increase of 13.07 in BLEU-1, as shown in Tab. 3. A key limitation noted by the baseline method was the issue of regression to the mean. This observation is quantified by the proposed total distance metric, which reveals that the progressive transformer generated motion that traversed only 25% of the ground truth distance. In contrast, all three teams produced more expressive outputs that more closely approximated the ground truth motion.

Team 2 outperformed the baseline on the hidden test set. However, on the PHOENIX14**T** test set, its performance decreased substantially, falling below the baseline. To understand this performance disparity between the two test sets, we analyzed the predicted sequences. We found that the average length of Team 2's predictions was 2.463 times longer than the ground truth on the PHOENIX14**T** test set, whereas on the hidden test set, it was close to 1. In comparison, Teams 1, 3, and the baseline exhibited average duration ratios of 1.438, 1.026, and 0.999, respectively. Just as spoken languages employ intonation, rhythm, and stress to convey nuance and meaning, sign languages utilize prosody. Thus, losing this information and producing longer sequences can be damaging to the performance. We hypothesize that Team 3's and the baseline's transformer-based architecture, ideally suited for sequence-to-sequence tasks, effectively captured this prosodic information.

Discrepancies emerged between certain pose-based and text-based metrics. Specifically, DTW-MJE tended to favor methods that produced less articulated or longer sequences. This observation underscores the importance of employing a diverse range of evaluation metrics when assessing generative models.

Across both test sets, Team 3 consistently produced the lowest Word Error Rate (WER). Notably, this approach incorporated BERT, a model pre-trained specifically on German. Team 1, which fine-tuned XLM-R, a multilingual

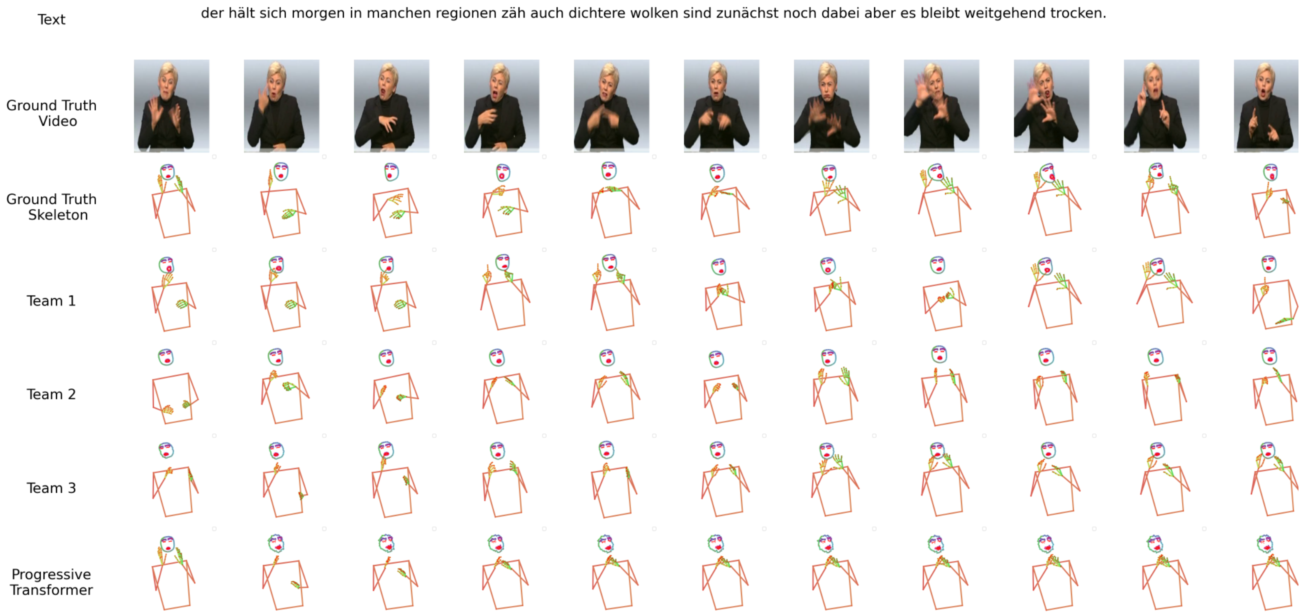
Text
der hält sich morgen in manchen regionen zäh auch dichtere wolken sind zunächst noch dabei aber es bleibt weitgehend trocken.
Ground Truth Video
Ground Truth Skeleton
Team 1
Team 2
Team 3
Progressive Transformer

Figure 6. Qualatative results of the top 3 performing teams on the PHOENIX14**T** test set. The first row shows the input spoken language sentence, the second and third rows show the ground truth video and skeleton, respectively, and the subsequent rows show the predicted skeleton poses from the top 3 teams. The final row shows the progressive transformer baseline.

model, demonstrated superior translation quality, as evidenced by higher BLEU scores. Given the low-resource nature of sign language, we hypothesize that the model's pre-existing linguistic knowledge significantly contributed to this performance. This suggests that spoken language resources can be effectively leveraged for SLP.

A correlation was observed between team rankings and model size; Team 1's model contained 13.7 times more parameters than Team 3's, and 2.8 times more than Team 2's.

Decomposing the WER metric into its constituent components revealed that each team predominantly exhibited errors within a distinct category. Specifically, Team 1 encountered the most replacement errors, Team 2 the most deletion errors, and Team 3 the most insertion errors. The hidden test set exhibited an average sentence length of eight words, with a range from three to fifteen words. A negative correlation was observed, on average, between sentence length and error rate across all three teams. This suggests a potential tendency towards over-translation, although this effect could originate from either the back-translation model or the teams' respective approaches, a distinction that is challenging to isolate. Overall, the words 'und,' 'regen,' and 'süden' were identified as the most frequent sources of error in the translations.

Qualitative results: Evaluation of the qualitative results supports the previous findings. The impact of the increased BLEU-1 score, noted for Team 1, is demonstrated in Fig. 6. This retrieval-based approach ensures that each generated sequence exhibits expressive signing. This observation is further quantified by our proposed total distance metric. As presented in Table 3, Team 1 was the only team to achieve a score surpassing one. Overall, the outputs from all teams preserved features of the ground truth sequences. However, further research is necessary to achieve the fluency and naturalness of native Deaf signers.

8. Conclusion

This paper summarizes the findings of the first SLP challenge. In the final test phase, six teams submitted solutions that outperformed the baseline approach. We presented the work of the top three teams. The top-performing team utilized a recognition model to curate a dictionary of signs, which was subsequently employed for production. We note that gloss annotations are a limited resource and therefore, would limit scaling this method to larger datasets. Teams 2 and 3 circumvented the requirement for this type of annotation and presented models that demonstrated competitive performance despite the absence of linguistic annotation.

Evaluating sign language is a challenging task that is compounded by the fact that data normalisations and representations are constantly evolving. We hope this work helps to establish a more consistent baseline for future SLP research. We implore future researchers to release their process features for other data sets, as without consistent inputs, comparison between approaches is meaningless.

Acknowledgments

This work was supported by the SNSF project 'SMILE II' (CRSII5 193686), the Innosuisse IICT Flagship (PFFS-21-47), EPSRC grant APP24554 (SignGPT-EP/Z535370/1), Google DeepMind and through funding from Google.org via the AI for Global Goals scheme. This work reflects only the author's views and the funders are not responsible for any use that may be made of the information it contains.

References

[1] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: a diffusion model for 3d sign language production from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1985–1995, 2024. 1, 3

[2] Andrew Bangham, SJ Cox, Ralph Elliott, John RW Glauert, Ian Marshall, Sanja Rankov, and Mark Wells. Virtual signing: Capture, animation, storage and transmission-an overview of the visicast project. In *IEE Seminar on speech and language processing for disabled and elderly people (Ref. No. 2000/025)*, pages 6–1. IET, 2000. 1, 2

[3] Jan Bungeroth and Hermann Ney. Statistical sign language translation. In *sign-lang@ LREC 2004*, pages 105–108. Citeseer, 2004. 2

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793, 2018. 3

[5] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 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*, pages 10023–10033, 2020. 4

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019. 2

[7] Sheng Chen, Qingshan Wang, and Qi Wang. Semantic-driven diffusion for sign language production with gloss-pose latent spaces alignment. *Computer Vision and Image Understanding*, 246:104050, 2024. 1, 2

[8] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. In *NeurIPS*, 2022. 5

[9] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019. 5

[10] Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. Tessa, a system to aid communication with deaf people. In *Proceedings of the fifth international ACM conference on Assistive technologies*, pages 205–212, 2002. 1, 2

[11] Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Lefebvre-Albaret. The dicta-sign wiki: Enabling web communication for the deaf. In *International Conference on Computers for Handicapped Persons*, pages 205–212. Springer, 2012.

[12] Oussama ElGhoul and Mohamed Jemni. Websign: A system to make and interpret signs using 3d avatars. In *Proceedings of the Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT), Dundee, UK*, 2011. 1, 2

[13] Sen Fang, Chunyu Sui, Xuedong Zhang, and Yapeng Tian. Signdiff: Learning diffusion models for american sign language production, 2023. 3

[14] Sylvie Gibet, François Lefebvre-Albaret, Ludovic Hamon, Rémi Brun, and Ahmed Turki. Interactive editing in french sign language dedicated to virtual signers: Requirements and challenges. *Universal Access in the Information Society*, 15: 525–539, 2016. 2

[15] John RW Glauert, Ralph Elliott, Stephen J Cox, Judy Tryggvason, and Mary Sheard. Vanessa–a system for communication between deaf and hearing people. *Technology and disability*, 18(4):207–216, 2006. 2

[16] Zhengsheng Guo, Zhiwei He, Wenxiang Jiao, Xing Wang, Rui Wang, Kehai Chen, Zhaopeng Tu, Yong Xu, and Min Zhang. Unsupervised sign language translation and generation, 2024. 3

[17] Jiayi He, Xu Wang, Ruobei Zhang, Shengeng Tang, Yaxiong Wang, and Lechao Cheng. Text-driven diffusion model for sign language production. *arXiv preprint arXiv:2503.15914*, 2025. 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 5

[19] Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. Towards fast and high-quality sign language production. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3172–3181, 2021. 2

[20] Eui Jun Hwang, Huije Lee, and Jong C Park. Autoregressive sign language production: A gloss-free approach with discrete representations. *arXiv preprint arXiv:2309.12179*, 2023. 2

[21] Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. Improving 3d pose estimation for sign language. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5, 2023. 3

[22] Jakub Kanis, Jiří Zahradil, Filip Jurčíček, and Luděk Müller. Czech-sign speech corpus for semantic based machine translation. In *International Conference on Text, Speech and Dialogue*, pages 613–620. Springer, 2006. 2

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–15, 2015. 6

[24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 4

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 3

[26] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 2

[27] Achraf Othman and Mohamed Jemni. Statistical sign language machine translation: from english written text to american sign language gloss, 2011. 2

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 4

[29] Anton Pelykh, Ozge Mercanoglu Sincan, and Richard Bowden. Giving a hand to diffusion models: a two-stage approach to improving conditional human image generation, 2024. 3

[30] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, 2015. Association for Computational Linguistics. 4

[31] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Adversarial training for multi-channel sign language production. In *British Machine Vision Virtual Conference*, 2020. 2

[32] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*, 2020. 3

[33] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 687–705. Springer, 2020. 1, 2, 4, 6, 7

[34] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Mixed signals: Sign language production via a mixture of motion primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[35] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[36] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5141–5151, 2022. 2

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 5

[38] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and R. Bowden. Sign language production using neural machine translation and generative adversarial networks. In *British Machine Vision Conference*, 2018. 2

[39] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *IJCV*, 128(4):891–908, 2020. 1

[40] Stephanie Stoll, Armin Mustafa, and Jean-Yves Guillemaut. There and back again: 3d sign language generation from text using back-translation. In *2022 International Conference on 3D Vision (3DV)*, pages 187–196. IEEE, 2022. 1, 3

[41] Rachel Sutton-Spence and Bencie Woll. *The linguistics of British Sign Language: an introduction*. Cambridge University Press, 1999. 1

[42] Shinichi Tamura and Shingo Kawasaki. Recognition of sign language motion images. *Pattern Recognition*, 1988. 1, 2

[43] Shengeng Tang, Jiayi He, Lechao Cheng, Jingjing Wu, Dan Guo, and Richang Hong. Discrete to continuous: Generating smooth transition poses from sign language observation. *arXiv preprint arXiv:2411.16810*, 2024. 2

[44] Shengeng Tang, Feng Xue, Jingjing Wu, Shuo Wang, and Richang Hong. Gloss-driven conditional diffusion models for sign language production. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2024. Just Accepted. 2

[45] Sümeyye Meryem Taşyürek, Tuğçe Kızıltepe, and Hacer Yalim Keles. Disentangle and regularize: Sign language production with articulator-based disentanglement and channel-aware regularization. *arXiv preprint arXiv:2504.06610*, 2025. 7

[46] Harry Walsh, Ben Saunders, and Richard Bowden. Changing the representation: Examining language representation for neural sign language production. In *LREC 2022 Workshop Language Resources and Evaluation Conference 24 June 2022*, page 117, 2022. 1

[47] Harry Walsh, Abolfazl Ravanshad, Mariam Rahmani, and Richard Bowden. A data-driven representation for sign language production. In *Proceedings of the 18th International Conference on Automatic Face and Gesture Recognition (FG 2024)*. Institute of Electrical and Electronics Engineers (IEEE), 2024. 1, 2

[48] Harry Walsh, Ben Saunders, and Richard Bowden. Select and reorder: A novel approach for neural sign language production. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14531–14542, 2024. 1

[49] Harry Walsh, Ben Saunders, and Richard Bowden. Sign stitching: A novel approach to sign language production. In *The 35th British Machine Vision Conference (BMVC)*, 2024. 1, 2

[50] Pan Xie, Qipeng Zhang, Zexian Li, Hao Tang, Yao Du, and Xiaohui Hu. Vector quantized diffusion model with codeunet for text-to-sign pose sequences generation. *arXiv preprint arXiv:2208.09141*, 2022. 1, 2

[51] Pan Xie, Taiyi Peng, Yao Du, and Qipeng Zhang. Sign language production with latent motion transformer, 2023. 3

[52] Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle

Guyon. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543, 2022. 3

[53] Jan Zelinka and Jakub Kanis. Neural sign language synthesis: Words are our glosses. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3384–3392, 2020. 2

[54] Ronglai Zuo, Fangyun Wei, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. A simple baseline for spoken language to sign language translation with 3d avatars. In *European Conference on Computer Vision*, pages 36–54. Springer, 2024. 1, 2, 3

[55] Inge Zwitserlood, Margriet Verlinden, Johan Ros, Sanny Van Der Schoot, and T Netherlands. Synthetic signing for the deaf: Esign. In *Proceedings of the conference and workshop on assistive technologies for vision and hearing impairment (CVHI)*, 2004. 1, 2