How much Data is Enough Data? A New Motion Capture Corpus for Probabilistic Sign Language Generation

Anna Klezovich[1], **Johanna Mesch**[1,2], **Gustav Eje Henter**[1], **Jonas Beskow**[1]
[1]Speech, Music and Hearing, KTH Royal Institute of Technology, [2]Stockholm University
Stockholm, Sweden,
johanna.mesch@su.se, {annkle, ghe, beskow}@kth.se

Abstract

We present a new 4.1 hours long high-quality motion capture sign language dataset for Swedish Sign Language — STS Mocap v1. The dataset consists of high quality multimodal data: body tracked with markers, fingers tracked with Manus Quantum Metagloves, face tracked with the iPhone LiveLink app in MetaHuman Animator mode, and corresponding textual sentence translation to spoken Swedish. With the help of this dataset, we show that four hours of motion capture data is enough for generative modeling of sign language conditioned on 2D pose. In comparison, training the same flow-matching model on only 30 minutes of this data, which is a common size for sign language motion capture datasets, shows a very apparent degradation in the quality of the synthesized data.

Keywords: motion capture dataset, multimodal dataset, sign language processing, motion generation

1. Introduction

Signed languages are used by 70 million people worldwide. And yet, in the recent revolution in generative AI, signed languages have to a large degree been left behind. Although sign language production (SLP) research has received increased interest, sign languages remain a low-resourced domain. SL video datasets are typically only in the range of tens to hundreds of hours, e.g., How2Sign ASL (79 hours) (Duarte et al., 2021) and STS-korpus (25 hours) (Mesch et al., 2012). A handful of large datasets of Internet and broadcast SL data have recently been released, such as YouTubeSL25 (Tanzer and Zhang, 2024) (3 000 hours) and BOBSL (1 500 hours) (Albanie et al., 2021). Such diverse large-scale datasets will be essential for AI models to learn linguistic variability for recognition tasks. SLP tasks, however, typically require well-structured data from a single signer. 2D video generation models based on GANs (Saunders et al., 2022) or diffusion models (Fang et al., 2025b) for a pose-to-video task can yield photorealistic results. However, this class of models is also known to introduce artifacts, since they lack explicit knowledge about 3D space and human anatomy. In general, video-based SLP approaches have not yet achieved the human-likeness that would lead these models to be used by the Deaf community or language learners (see, for example, state-of-the-art generated videos from the recent Fang et al. (2025a) SignLLM paper).

Sign language is inherently three-dimensional, which makes 3D avatars a compelling alternative for generating human motion. When trained on 3D motion capture data, current generative motion models can reproduce high-quality human motion across different domains, such as locomotion, fighting, dance, and co-speech gesture. For example, one study (Alexanderson et al., 2023) built motion generation experiments using just 2–4 hours of data for co-speech generation, 3.5 hours for dance, and in the most data-intensive locomotion generation experiment, up to 18.8 hours. As a result, their diffusion model showed a higher human-likeness than multiple strong baselines across three domains of human motion data on as little as 2 hours of training data. Moreover, one of the most-used co-speech gesture datasets featured in a number of generative AI papers (Mehta et al. (2024), Alexanderson et al. (2023), Alexanderson et al. (2020), etc.) is the 4-hour Trinity Speech-Gesture Dataset 2 (TSG2) (Ginosar et al. (2019), Ferstl and McDonnell (2018), (Ferstl and McDonnell, 2021)).

Consequently, it was reasonable to assume that as little as 4 hours of high-quality 3D data could also produce good results for sign language generation. To pursue this idea, we collected 4.1 hours of high-quality motion capture data for Swedish Sign Language (STS), trained generative models of sign language, and evaluated the results. With this paper, we are also releasing the code[1] and the first version of the dataset[2].

Our main contributions are:

- Presenting a high-quality motion capture dataset for Swedish Sign Language, including detailed hand and face motion.

[1]STS-mocap-dataset-v1-sample provides preprocessing and synthesis scripts, a sample of motion files, a sample of corresponding renders both for original files and synthesized.

[2]STS-Mocap-v1 dataset on Huggingface.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9549–9558
11–16 May 2026.

Dataset	Language	Duration	Body	Fingers	Face	Annotation
TSG2 (2019)		4h	53 markers total for both body and fingers		no	Aligned audio
BEAT2 (2024)		60h	MoSh mesh on 27 markers	MC-Mesh on 24(×2)	FLAME mesh on 51 BS^a	Aligned audio, text
LSF-SHELVES	LSF	$>$0.5h	Kinect Azure, 32 joints	no	no	no
LSF-ANIMAL	LSF	$\approx$ 1h	55 markers	26(×2) markers	16 markers	Glosses, Phonology
STK LSF	LSF	$\approx$ 1h	35 markers	20(×2) markers	51 BS^a & 40 markers	no
Deep JSLC	JSL	unknown	31 markers	24(×2) markers	44 makers	SignWriting, Glosses
CUNY ASL	ASL	$\approx$ 3.5h	Upper body markers	Immersion CyberGloves	eye-tracker	Glosses, Syntax, Non-manuals
STS Mocap v1	STS	4.1h	50 markers	Manus Gloves	MHA LiveLink	Sentences
MC-TRISLAN	CSE	18h	46	3 markers^b	7 markers	Glosses, Handshapes

^{a} – ARKit blendshapes; ^{b} – infilled with the nearest handpose from a separately recorded set of handposes;

Table 1: A table comparing the size of this dataset to other SL mocap datasets and co-speech gesture datasets.

- Demonstrating that the dataset is sufficiently large for a simple generative modeling task.
- Introducing a generative probabilistic model based on conditional flow-matching for 3D sign language motion generation.

The remainder of the paper is structured as follows: Section 2 presents an overview and comparison of sign language motion capture datasets; Section 3 describes the data collection procedure; Section 4 provides statistics on the dataset; Section 5 describes our sign language generation experiments; and Section 6 presents the results.

2. Background

Several sign language motion capture datasets have been published across a number of sign languages. In this section, we give an overview of the six most prominent among them. Table 1 compares these six datasets with our own, along with two co-speech gesture motion capture datasets for reference. The table shows the duration of the recordings, how the body, fingers, and face were recorded, and the types of annotation they include.

There are at least three motion capture datasets for French Sign Language (LSF). The LSF-ANIMAL dataset (Naert et al., 2020) contains around 1 hour of data in the domain of animal names and properties. It consists of recordings of individual signs for animals, recordings of all the phonological handshapes of LSF, and narratives with descriptions of said animals. LSF-ANIMAL also offers a detailed human evaluation pipeline for sign language motion capture. LSF-SHELVES (Mertz et al., 2022) is not a standard markers-based motion capture dataset; it was captured with Kinect Azure and recorded with one infrared camera tracking only body motion. However, it is still of interest because it specifically covers spatial constructions and referencing. The SignToKids LSF dataset (Reverdy et al., 2024) has approximately 1 hour of signing data and focuses on one domain – tales for schoolchildren. SignToKids is recorded with only one signer and is also very multimodal. The face recordings had markers, blendshapes, and separate gaze tracking, however, the authors reported that they had to do a lot of manual editing and manual occlusions infilling.

CUNY ASL (Lu and Huenerfauth, 2012) provides approximately 3.5 hours of motion capture data focusing on the upper body, recorded with 8 different signers. Face data is not captured, however, the authors captured eye gaze. The domain is not restricted; they used nine types of prompts for the signer in a motion capture costume, one of them was in the news domain. The data is advertised as available upon request via email.

The Japanese SL dataset by Brock and Nakadai (2018) consists of 10 000 translated sentences recorded with a single CODA signer. The duration of the dataset is not reported. All of the data was captured with physical markers, including 44 markers on the face. Interestingly, the authors experimented with sequence-to-sequence sign language recognition networks to demonstrate that the data was useful for model training, but achieved only $\approx$ 40% accuracies for their sign-to-gloss task.

The largest sign language motion capture dataset created to date is the Czech SL (CSE) dataset ((Krňoul et al., 2023); (Jedlička et al., 2022)), MC-TRISLAN. MC-TRISLAN was recorded with 59 body markers, including 3 markers on each hand and only 7 markers on the face. The authors also used 21 markers to record detailed right hand transitions from a neutral hand pose to a set of hand poses. The handshapes recorded were specific to CSE phonology. In post-process, hand pose was algorithmically infilled with the nearest hand pose from these detailed recordings. The dataset consists of two types of stories: weather forecasts and zoo tours, and is recorded with several signers.

Figure 1: Motion capture set up.

Figure 2: Example frames from our dataset visualized with different MetaHuman Avatar presets (Left to right: Oskar (male, average hight, medium weight), Neema (female, tall, underweight), Jesse (male, tall, medium weight)).

3. Methods

In this section, we describe our motion capture recording pipeline: it is multimodal, it offers state-of-the-art face capture, and it does not require any manual infilling. The only manual work required is to fix marker swaps and apply pattern interpolation where needed using the Motive software (NaturalPoint, Inc., 2024), which is intuitive and does not require 3D engineering expertise. The dataset is recorded simultaneously for all types of motion capture and relies fully on automated procedures.

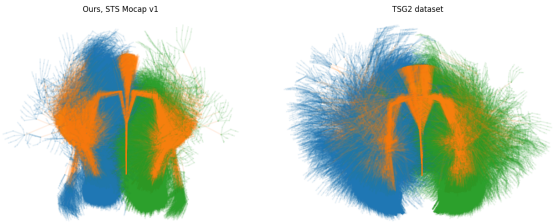

Figure 3: A 2D projection of every 20th frame of our data Vs. TSG2 data. Hips are centered at (0, 0); TSG2 data is retargeted to our skeleton.

3.1. Dataset Contents

Since our dataset is primarily designed for training of generative models, we restricted the domain and the number of signers to one. With a single signer, the model does not need to disentangle the differences between the signers from the semantically relevant information (similarly to TTS (Liu et al., 2022)). At the same time it also somewhat reduces the potential vocabulary size and the overall diversity of the signing, which makes the dataset less linguistically diverse, but much more manageable for machine learning tasks.

The data was recorded with a single CODA signer, with a second deaf signer present for consultation for most of the recording sessions. To avoid priming the signer to switch to signed Swedish, the signer was first asked to read a sentence and think of the best way to translate it. Then, for the actual recording, the sentence would disappear so that the signer translated it from memory. After the recording of a sentence ended, the signer had the opportunity to read the same sentence one more time if needed. Multiple takes were allowed. In addition, the second signer could give suggestions on translations to the main signer.

The main part of the dataset (3 hours 47 minutes) consists of news articles from *8sidor* (MTM and Hillblom, 2025). These texts are written in *easy Swedish*, meaning they avoid long, complex sentences and complicated words. They were translated by the signer from Swedish to STS sentence by sentence.

The remaining 22 minutes of the data (8% of the dataset) is a trial dataset that we used to establish our recording pipeline. It includes 39 recordings of signs and their uses in sentences; both a sign and a sentence are repeated twice with a pause in between. These recordings were made with the same CODA signer and in the same setup, the only difference being the content recorded.

3.2. Motion Capture Setup

Figure 1 depicts our motion capture setup in action, with OptiTrack cameras around the room and a signer in a motion capture costume sitting in the middle in front of an iPhone camera on a tripod. Text and video prompts appeared on the TV screen in front of the signer, and were stored with the motion on each take.

Our STS Mocap v1 dataset has three types of motion capture data: body, fingers, and face. The body was tracked with 25 OptiTrack Prime cameras and 50 passive markers. One of the cameras was located directly above the signer to capture the top view and hand motions close to the body. Fingers were tracked with Manus Quantum Metagloves. Body and fingers were recorded at 120 Hz. Body and fingers were solved to a skeleton

together in the Motive software (NaturalPoint, Inc., 2024). The minimal manual postprocessing was performed in Motive. It involved fixing the occasional mislabeling of markers and running linear or pattern interpolation on occluded segments, where applicable.

The face was tracked at 30 Hz with the iPhone LiveLink app using MetaHuman Animator mode. MetaHuman Animator mode was chosen over ARKit because it offers superior face reconstruction accuracy and a more detailed animation model (173 controls over 51 blendshapes). MetaHuman Animator also yields better intelligibility for sign language than ARKit (Klezovich et al., 2025). Then, the face depth data was processed in Unreal Engine to a control rig. The iPhone LiveLink face data is designed to be compatible with a MetaHuman avatar and its face control rig. Example frames visualized with different MetaHuman avatars are depicted in Figure 2, (Epic Games, 2025). Our dataset also includes close-captured videos of the face, so that any video-based face tracking method could be applied to drive other face rigs. For example, ARKit face blendshapes could be extracted with the help of MediaPipe (Google Research, 2019).

Naturally, this setup was somewhat limited in that the signer had to stay in the frame of the iPhone camera, which made the signing more restricted than the natural signing. The signer had to consciously move their body and their head slightly less. When the signer did move their body or head it often lead to a couple of frames being occluded in the face depth data. The occluded frames were linearly interpolated in the Unreal Engine, which as we have showed previously (Klezovich et al., 2025) does not affect the intelligibility of the signing avatar. A headmount did not work for us, because the iPhone was too heavy for it, and that would introduce a slight tremor to the recorded face depth data, making it unusable.

4. Dataset Statistics

We collected 4 hours and 9 minutes of data in total. The dataset consists of 1 772 sentence recordings with repeating sentences, 1 408 of which are unique sentences. The corresponding Swedish sentences were processed with the *'sv core news sm'* *spacy* pipeline (Honnibal et al., 2020), (Explosion, 2024). The vocabulary size of this set of sentences is 1 983 lemmas. The duration of the dataset when counting only the last take of each unique sentence is 3 hours and 5 minutes.

Compared to the Trinity Speech Gesture 2 dataset (Ferstl and McDonnell, 2021), our raw BVH files have a higher percentage of zero motions. While the statistics for the body are much more similar (1% in our data and 0.06% in the TSG2 data), 35% of finger data in our dataset has zero motions in it. Essentially, the fingers, which were recorded with Manus Gloves, are at a different frame rate than the body, because the median length of a zero-motion segment is 1 frame for all finger joints and the mean length is 1.6 for left hand finger joints and 1.8 for right hand finger joints. This should be accounted for when modeling on this dataset, either with interpolation or downsampling of the motion.

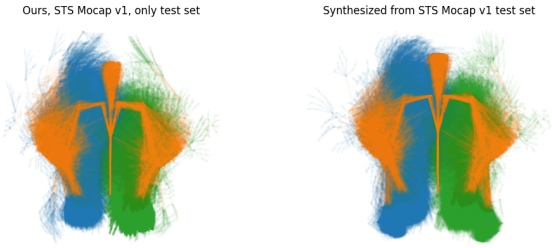

Figure 4: A 2D projection of every 20th frame of our test set data Vs. data synthesized from it.

To give a qualitative metric for the diversity of collected poses, we created an overlay of poses (taking each 20th frame) for our dataset and for the Trinity Speech Gesture 2 dataset (see Fig. 3). It shows the poses in our dataset are more localized and closer to the body, which is likely an effect of the sign language being more structured than spontaneous gesture, as well as the fact that the signer was sitting on a chair and was instructed to try to stay in the frame of the iPhone camera capturing their face, while in the TSG2 dataset the actor was allowed to move more.

5. Experiments

In order to demonstrate the possible use case for this dataset, we trained optimal transport conditional flow-matching models (OT-CFM) following co-speech gesture generation papers (Mehta et al., 2024), (Liu et al., 2025), and (Alexanderson et al., 2020). (Mehta et al., 2024) and (Alexanderson et al., 2020) are based on the TSG2 dataset (Ginosar et al., 2019), which is similar in size to our STS Mocap v1 dataset. Hence, the expectation is that there is enough data for generative modeling.

Conditional flow-matching models introduced by Lipman et al. (2023) learn to predict a probability path from noise distribution to the data distribution. In contrast to diffusion-based methods, Optimal Transport (OT) displacement interpolation defines the conditional probability paths between the noise and the target vector fields, instead of defining diffusion paths, which means that the model is faster at synthesizing the data without losing quality. On each training step, the OT-CFM model samples a random time step between noise and data distribu-

tion $t \in [0, 1]$ and takes the interpolation between data and noise for this time step as an input to the network. The training loss is a mean squared error between the velocity field predicted by the network and the target velocity field, or as is formalized in (Lipman et al., 2024):

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t, X_t, X_1}\left[\left\| u_t^\theta(X_t) - u_t(X_t \mid X_1)\right\|^2\right],$$

where time step $t \sim \mathcal{U}[0, 1]$; X_t is an intermediate data sample representing a linear combination of X_0 and X_1 at t; $u_t(X_t \mid X_1)$ is a **target** conditional velocity field, and $u_t^\theta(X_t)$ is an **input** velocity field and θ are learnable parameters of the network.

5.1. Training Objective

OT-CFMs can be used in a number of generative tasks, for example in a text-to-speech task (Mehta et al., 2023) or in a text to speech and gesture task (Mehta et al., 2024). The current STS Mocap v1 dataset is only coarsely annotated, meaning that sentences are not time aligned with signing and there is no annotation for glosses so far. In this paper, we experiment with the task of 3D lifting to demonstrate that this dataset is suitable for generative modeling. The model is trained to predict 3D motion sequences based on 2D poses.

To investigate the impact of dataset size to generative modeling in the particular task of predicting 3D sequence from 2D input sequence, we also trained a model on just half an hour of our data (12% of the dataset).

During training, the model takes 2D projected motion sequences of the 3D motion data from the training set as an input and the corresponding 3D data as a target. For the training, the data is windowed into shorter motion sequences of a fixed length. During inference, the model takes input 2D sequences of any length and predicts the output 3D sequences of the same length.

5.2. Architecture

The architecture of the OT-CFM model is, for the most part, adapted from the Matcha-TTSG model introduced in (Mehta et al., 2024). It is a U-net network built with 1D convolutional Resnet layers followed by transformer blocks. The main difference in our approach from the Matcha-TTSG (Mehta et al., 2024) model is that we train three separate streams, having body data, left hand data, and right hand data flow through separate networks and separate optimizers, but on the same batches of data and same time frames at a time. Each batch of data is split into three before the training step, and there are three optimizers doing backwards step based on three separate losses. In the Matcha-TTSG model, speech and gesture generation is unified by having the data concatenated before passing it as a target vector field into the flow-matching model. Our initial experiments with the unified streams for body, right hand, and left hand on 1D convolutional U-net produced some jitter in the synthesis, so we opted to have 3 networks to make it easier for the model to learn. The non-unified approach also makes it easier to iteratively add streams to the model, and in the future, to add the face generation stream.

Before training, the data was preprocessed with the help of *pymo* python library (Alemi, 2019). The data was downsampled with a factor of 4, from 120 to 30 frames per second. Then it was centered around the hips root joint. Only the upper body joints were selected for training. The data was also mirrored to increase the size and variability of the training set. The near-zero velocity frames were cropped from the beginning and the end of each sentence recording. The Euler angles were then transformed into exponential maps in radians and normalized over the training dataset. Recent work by Fauré et al. (2025) argues for the benefits of using quaternions, which we plan to include in the future.

The input data had a fixed sequence length extracted with a sliding window with 1/2 overlap. We experimented with different input sequence lengths, between 10 and 120 frames. The model was trained with early stopping based on the validation loss aggregated across three streams. For evaluation, we take the best model checkpoint before overfitting.

6. Results

OT-CFM models trained with 2D poses as conditions were evaluated quantitatively by calculating the $RMSE$ error between the 2D projections of synthesized 3D sequences and the corresponding input 2D poses. The input 2D poses come from projections of the 3D test set. The test set consists of 276 sentence recordings. In this evaluation all of the data was synthesized with 22 steps of ODE solver.

The results are reported in Table 2 for both *full* models trained on the whole dataset and *short* models trained only on 12% or ≈30 minutes of data. We trained models with three sequence lengths for comparison: 10, 28, and 120 frames (the data is at 30 fps). The errors are reported separately for the body, the right hand, and the left hand.

$RMSE$ error was standardized by variation to understand the scale of the difference between the calculated values. $RMSE_{variation}$ is calculated by taking $RMSE$ between the 2D condition motion sequence and the 2D mean pose for our training data. In Table 2 it is reported separately for models trained on the whole dataset and models trained

		$RMSE_{error}$ **standardized by variation,** CI		
Model	**Seq length**	**Body**	**R Hand**	**L Hand**
full	10	0.13 (0.12, 0.14)	0.17 (0.16, 0.17)	0.19 (0.18, 0.20)
full	28	**0.11 (0.10, 0.12)**	**0.16 (0.15, 0.16)**	**0.16 (0.16, 0.17)**
full	120	0.14 (0.14, 0.14)	0.20 (0.20, 0.21)	0.20 (0.20, 0.21)
short	10	**0.11 (0.10, 0.11)**	0.19 (0.18, 0.19)	0.19 (0.18, 0.19)
short	28	0.13 (0.13, 0.14)	0.21 (0.21, 0.22)	0.22 (0.21, 0.23)
short	120	0.27 (0.27, 0.28)	0.36 (0.35, 0.37)	0.36 (0.35, 0.37)
		$RMSE_{variation}, CI$		
full		5.20 (5.13, 5.29)	14.18 (13.90, 14.33)	14.19 (13.90, 14.34)
short		5.26 (5.19, 5.34)	14.23 (13.93, 14.41)	14.23 (13.93, 14.42)

Table 2: Standardized RMSEs with 95% CIs, aka. a value < 1 means error is lower than variation in the data.

Data			**Jerk**	**Acceleration**
test data			6036	330
training data			5837	319
	seq len	**ODE steps**		
full model, synthesized	10	22	8500	418
full model, synthesized	28	15	**6545**	**345**
full model, synthesized	28	22	6615	347
full model, synthesized	28	34	6712	351
full model, synthesized	28	50	6749	352
full model, synthesized	120	22	12253	539
short model, synthesized	10	22	11112	501
short model, synthesized	28	22	14744	611
short model, synthesized	120	22	52315	1945

Table 3: Average acceleration magnitude and average jerk magnitude for original data compared to data synthesized with different models at different number of ODE steps.

on half an hour of data, because the mean poses are extracted from the respective training sets.

First, $RMSE$ error is calculated between the 2D condition data sequence and the model's output 2D projection sequence. Second, it is divided by the respective $RMSE_{variation}$. The standardized $RMSE$ error shows the size of the error compared to the size of the variation. The smaller this ratio, the better.

All values reported in Table 2 are medians with confidence intervals that are estimated with the help of nonparametric bootstrap. This way, we avoid making statistical assumptions about the underlying distribution of errors. The data is resampled with 10 000 bootstrap samples. Confidence intervals calculated at $\alpha = 0.05$.

Table 2 shows that full models perform better than short models on hands data. Standardized $RMSE$ for the best full model (28 sequence length) for the right hand is 0.16, while for the short model trained on the same sequence length it is 0.21. The same is true for the left hand. The lowest standardized $RMSE$ for the full model is 0.16 while for the same short model it is 0.22. The lowest standardized $RMSE$ for all short models is 0.19, both for the right and the left hand data. Comparing models with respect to the body data standardized $RMSE$ does not give such a consistent result. In this case, the short model has the same lowest standardized $RMSE$ as the full model – 0.11. The short model slightly outperforms the full model on the body data when trained on a sequence length of 10 frames. But the full model leads on its performance on the hands data.

As an additional test, we compared the jerk and acceleration metrics on the original data Vs. the synthesized data, following Kucherenko et al. (2024). Although these metrics are known to not correlate with human likeness (Kucherenko et al., 2024), in this case, they show how similar the model outputs are to the ground truth test data. Table 3 shows average acceleration and average jerk magnitudes for the test data used for conditioning (aka. the 3D motion data that was projected to 2D for conditioning), for training data, and for the full models and the short models trained on three types of sequence lengths. As an ablation, we also synthesized data from the best full model (sequence

Model, seq length, ODE steps	Body	R Hand	L Hand
	Euclidean dist (cm), CI		
full, 28, 22	1.24 (1.21, 1.28)	3.93 (3.85, 4.01)	3.93 (3.85, 4.01)
short, 28, 22	1.50 (1.44, 1.53)	5.35 (5.20, 5.46)	5.35 (5.20, 5.46)
	Euclidean dist standardized by variation, CI		
full, 28, 22	0.21 (0.20, 0.21)	0.20 (0.20, 0.21)	0.20 (0.20, 0.21)
short, 28, 22	0.24 (0.23, 0.25)	0.27 (0.26, 0.28)	0.27 (0.26, 0.28)
	$RMSE_{variation}$ (cm), CI		
full	5.99 (5.94, 6.04)	20.17 (19.95, 20.52)	20.18 (19.95, 20.53)
short	6.11 (6.07, 6.16)	20.48 (20.08, 20.86)	20.48 (20.07, 20.87)

Table 4: Top half: bootstrapped median Euclidean distances between the **3D** ground truth sequences corresponding to input conditions and output 3D synthesized sequences with 95% CIs; Bottom half: standardized Euclidean distances with 95% CIs;, aka. < 1 means error is lower than variation in the data.

length 28) with different numbers of ODE solver steps to make sure that the difference from the test set metrics actually comes from the model and not from the lack of the synthesis step. This table shows that the data synthesized with the different number of ODE steps results in similar metrics with a negligible difference. The best full model trained on 28 frames long sequences shows average acceleration magnitude and jerk magnitude that are very similar to the test data. If we average between the four statistics we got for a the different number of ODE steps, we can show that the full model produces metrics that are only 6% higher for acceleration and 10% higher for jerk. The short model performs much worse. The best short model is about 2 times worse than the ground truth test data, or, more specifically, the acceleration is 52% higher and the jerk is 84% higher.

Since we synthesize the data from projections of the test set, it means we have the ground truth 3D data. We can compare our 3D data from the test set with the 3D model's predictions for full models and short models with the same settings, aka. the sequence length of 28 for training and 22 ODE steps for synthesis. In a similar fashion as before, we measure Euclidean distances between the points in the 3D space for each joint, averaged within each sequence. 3D positions inferred from the BVH data are in centimeters. The distances reported in the table are bootstrap median values with confidence intervals at $\alpha = 0.05$ and 10 000 bootstrap samples. The standardized Euclidean distance is calculated with dividing Euclidean distance by variation, so when this ratio is < 1, the error is lower than the variation in the data. The results for comparing the output 3D sequences with the ground truth sequences show that the full model has lower error than the short model, trained on 12% of our data, for body and both hands.

For qualitative evaluation, we plot a poses overlay for the best model trained with sequence length 28 and synthesized with 22 ODE steps compared to a poses overlay for test set data that was used to condition the synthesis (Fig. 4). In addition, we provide a sample of generated BVH files and their respective conditions in our repository and their skeleton renders for an easy visual comparison[3].

7. Conclusion

In this paper, we address the issue that sign languages are a low-resource domain for ML training. Although there are many video-based sign language datasets, models trained on videos have not reached the levels of human-likeness needed for practical application. This is because of the domain gap between 2D and 3D data — video data leads to a loss of some depth information, which introduces noise.

To address this need for high-quality motion capture data for sign languages, we have collected a novel motion capture dataset, STS Mocap v1. Our dataset consists of 4.1 hours of sentence recordings. Recordings were made with a single signer and in one domain (simplified news) in order to avoid inter-signer variability and keep vocabulary size limited.

We demonstrate that 4 hours of sign language motion capture data is sufficient for generative modeling, at least for a 2D to 3D task with a flow matching model. To show this, we trained several OT-CFM models with the objective of lifting 2D poses to 3D joint angle data. The model produces smooth and visually accurate motion sequences. We evaluated the model quantitatively by comparing the input 2D condition poses with the output 2D projection poses from 3D motion sequences on the test set. The model produces an RMSE that is much lower than the unit variance, which means that the

[3]A link to the folder with side-by-side skeleton renders: STS-mocap-dataset-v1-sample

model is learning reasonable 3D data reconstructions.

We also compare the model trained on the full dataset to the model trained on half an hour of data or 12% of the same dataset. As a result, the model trained on less data demonstrates much worse jerk and acceleration metrics, around 2 times worse than the full model and about 2.6 times worse than the training data. The full model also shows better standardized RMSE on all three streams of data.

This proves that as little as 4 hours of sign language motion capture data is enough for this generative modeling task, while taking half an hour of the same data very noticeably degrades the model quality.

8. Discussion and Limitations

This dataset has many other potential use cases and experiments which could be run. For example, we have not yet utilized the recorded face animation data. In future, it would be interesting to see whether this model is able to deal with face animation generation based on just ≈ 4 hours of data. Face animation has much higher dimensionality (173 controls) than body or hands data, so it might be harder for this model to learn it without dimensionality reduction. In addition to that, in this paper we compared only two sizes of data for training: 30 minutes and 4 hours. It would be also potentially interesting to conduct an ablation study with different dataset sizes to determine whether or not there is an optimal dataset size for this type of model.

We also included sentence level annotations in this dataset. They could be automatically frame aligned and even sparsely annotated for glosses using the STS corpus (Mesch et al., 2012). After that this dataset could also be tested for a more complex task – a text-to-sign task. However, as of now, we do not know how well the current model and this dataset would generalize to the text-to-sign task. This kind of experiment could provide an evaluation of the linguistic quality of this data.

The model architecture itself could also be adjusted. In the current model, there are three independent streams, one for the body and one for each of the hands. Each stream has a different level of the data complexity which is both a motivation for at least somewhat separating them and a potential issue with the current training setup. Because of the different data complexity the stream for the body trains faster than the streams for the hands. Either streams for the hands can end up being undertrained or stream for the body can end up being overfitted. In addition, using three streams could potentially hinder the capture of inter-stream dependencies that are important for sign language. Currently, the only thing that connects the streams in the model is that the model sees body and hands from the same data sample at each step. Therefore, it could be beneficial for model training to fuse the streams in some way, for example using cross-attention. We also limit this study only to one flow-matching model, because it was shown to work for similar sized datasets in the domains of TTS and co-speech gesture generation (Mehta et al. (2023); Mehta et al. (2024)). There are sign language generation studies, such as SignDiff by (Fang et al., 2025b), where the authors trained a diffusion model, but on a much larger dataset – 79 hours of data, How2Sign dataset (Duarte et al., 2021). It would be very interesting to compare the diffusion models with the flow-matching models on our relatively small dataset.

Even though we have shown that the amount of data collected is sufficient for some generative modeling tasks, we plan to record 6 more hours of data and release it as the next version of this dataset, where we will include other domains besides news and possibly more signers, and use it for more challenging modeling tasks as part of text-to-signing avatar SLP pipelines.

9. Acknowledgements

This dataset collection was done within the KTH SignBot project funded by Vetenskapsrådet (grant nr. 2023-04548). We thank everyone who participated in the creation of this dataset and especially our main signer who tirelessly worked so many hours with us in the motion capture studio.

10. Bibliographical References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset. *arXiv*.

Omid Alemi. 2019. PyMO: Motion capture library. Accessed: 2025-10-14.

Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. *ACM Trans. Graph.*, 42(4).

Simon Alexanderson, Éva Székely, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Generating Coherent Spontaneous Speech and Gesture from Text. In *Proceedings of the 20th*

ACM International Conference on Intelligent Virtual Agents, IVA '20, New York, NY, USA. Association for Computing Machinery.

Heike Brock and Kazuhiro Nakadai. 2018. Deep JSLC: A multimodal corpus collection for data-driven generation of Japanese Sign Language expressions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Epic Games. 2025. Metahuman creator overview. Accessed: 2025-06-05.

Explosion. 2024. sv_core_news_sm: Swedish core pipeline for spaCy. Components: tok2vec, tagger, morphologizer, parser, lemmatizer, senter, ner. License: CC BY-SA 4.0. Sources: UD Swedish Talbanken v2.8; Stockholm–Umeå Corpus v3.0.

Sen Fang, Chen Chen, Lei Wang, Ce Zheng, Chunyu Sui, and Yapeng Tian. 2025a. SignLLM: Sign Language Production Large Language Models.

Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Yapeng Tian, and Chen Chen. 2025b. Signdiff: Diffusion model for american sign language production. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–11. IEEE.

Guilhem Fauré, Mostafa Sadeghi, Sam Bigeard, and Slim Ouni. 2025. Towards skeletal and signer noise reduction in sign language production via quaternion-based pose encoding and contrastive learning. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA '18, page 93–98, New York, NY, USA. Association for Computing Machinery.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *CoRR*, pages 3492–3501.

Google Research. 2019. MediaPipe.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Anna Klezovich, Johanna Mesch, and Jonas Beskow. 2025. Motion capture driven avatars for Swedish Sign language. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Zdeněk Krňoul, Pavel Jedlička, Miloš Železný, and Luděk Müller. 2023. Motion capture 3D Sign Language resources. In Georg Rehm, editor, *European Language Grid*, Cognitive Technologies, pages 307—-312. Springer, Cham.

Taras Kucherenko, Piete Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2024. Evaluating gesture generation in a large-scale open challenge: The genea challenge 2022. *ACM Transactions on Graphics*, 43(3):1–28.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow matching for generative modeling.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. 2024. Flow matching guide and code.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. 2024. EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1144–1154.

Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. 2025. GestureLSM: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10929–10939.

Songxiang Liu, Dan Su, and Dong Yu. 2022. Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans. *arXiv preprint arXiv:2201.11972*.

Pengfei Lu and Matt Huenerfauth. 2012. CUNY American Sign Language motion-capture corpus: First release. In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 109–116, Istanbul, Turkey. European Language Resources Association (ELRA).

Shivam Mehta, Ruibo Tu, Simon Alexanderson, Jonas Beskow, Eva Szekely, and Gustav Henter. 2024. Unified speech and gesture synthesis using flow matching. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024)*, pages 8220–8224.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2023. Matcha-TTS: A fast TTS architecture with conditional flow matching. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11341–11345.

Clémence Mertz, Vincent Barreaud, Thibaut Le Naour, Damien Lolive, and Sylvie Gibet. 2022. A low-cost motion capture corpus in French Sign Language for interpreting iconicity and spatial referencing mechanisms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 2488–2497, Marseille, France. European Language Resources Association.

Johanna Mesch, Lars Wallin, and Thomas Björkstrand. 2012. Sign language resources in Sweden: Dictionary and corpus. In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 127–130, Istanbul, Turkey. European Language Resources Association (ELRA).

Myndigheten MTM and Marie Hillblom. 2025. 8 Sidor: Nyheter på lätt svenska.

Lucie Naert, Caroline Larboulette, and Sylvie Gibet. 2020. LSF-ANIMAL: A motion capture corpus in French Sign Language designed for the animation of signing avatars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6008–6017, Marseille, France. European Language Resources Association.

NaturalPoint, Inc. 2024. Motive: Optical motion capture software.

Clément Reverdy, Sylvie Gibet, and Thibaut Le Naour. 2024. STK LSF: A motion capture dataset in LSF for SignToKids. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 315–322, Torino, Italia. ELRA and ICCL.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5131–5141, Los Alamitos, CA, USA. IEEE Computer Society.

Garrett Tanzer and Biao Zhang. 2024. YouTube-SL-25: A large-scale, open-domain multilingual Sign Language parallel corpus. *ArXiv*, abs/2407.11144.

11. Language Resource References

Amanda Duarte and Shruti Palaskar and Lucas Ventura and Deepti Ghadiyaram and Kenneth DeHaan and Florian Metze and Jordi Torres and Xavier Giro-i-Nieto. 2021. *How2Sign Dataset*. distributed via ELRA: ELRA-Id ELRA-S0416, ISLRN 583-408-694-292-6.

Ferstl, Ylva and McDonnell, Rachel. 2021. *Trinity Speech-Gesture dataset*. European Language Grid. PID http://live.european-language-grid.eu/cpid/HoZg9Qd7nYNDdUNsaUzyNF. [Dataset (Audio and Text corpus)].

Jedlička, Pavel and Krňoul, Zdeněk and Zelezny, Milos and Muller, Ludek. 2022. *MC-TRISLAN: A Large 3D Motion Capture Sign Language Data-set*. European Language Resources Association. PID https://live.european-language-grid.eu/catalogue/project/8179.