# SignLLM: Sign Language Production Large Language Models

Sen Fang[1], Chen Chen[2†], Lei Wang[3,4], Ce Zheng[5], Chunyu Sui[6], Yapeng Tian[7]
[1] Rutgers University, [2] University of Central Florida, [3] Australian National University, [4] Data61/CSIRO
[5] Carnegie Mellon University, [6] Columbia University, [7] The University of Texas at Dallas
†: Corresponding Author. Project Page: https://signllm.github.io

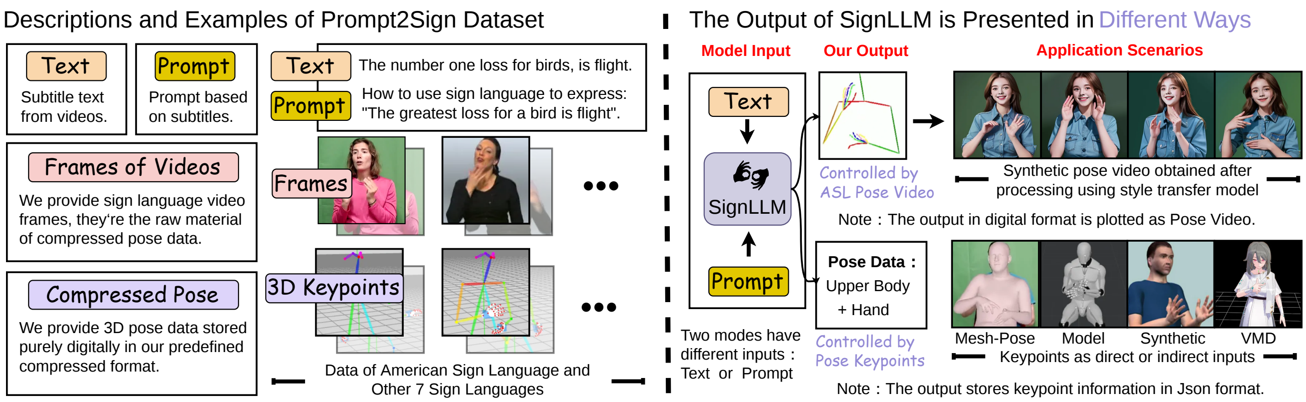


Figure 1. **Overview:** (Left) Major components (*e.g.*, Text, Prompt, Compressed Pose, *etc.*) of our PROMPT2SIGN dataset. Compressed Pose is reprocessed pose data that is suitable for training, we use public sign language videos to produce compressed pose data in our predefined format; (Right) Our proposed SIGNLLM aims to generate sign language poses for digital human or avatar generation [6, 12, 105, 110].

## Abstract

*In this paper, we propose* **SIGNLLM***, a multilingual Sign Language Production (SLP) large language model, which includes two novel multilingual SLP modes MLSF and Prompt2LangGloss that allow sign language gestures generation from query texts input and question-style prompts input respectively. Both modes can use a new RL loss based on reinforcement learning and a new RL module named Priority Learning Channel. These RL components can accelerate the training by enhancing the model's capability to sample high-quality data. To train* SIGNLLM*, we introduce* **PROMPT2SIGN***, a comprehensive multilingual sign language dataset, which builds from public data, including American Sign Language (ASL) and seven others. This dataset standardizes information by extracting pose information from sign language videos into a unified compressed format. We extensively evaluate* SIGNLLM*, demonstrating that our model achieves state-of-the-art performance on SLP tasks across eight sign languages.*

## 1. Introduction

Sign Language Production (SLP) aims to synthesize human-like sign avatars from text inputs. Deep learning-based SLP approaches [38, 71–74] typically involve sequential steps from text to gloss (*i.e.*, a type of textual vocabulary representing gestures or postures), gloss to pose [10, 52], and finally rendering pose videos into more engaging human-like avatar videos. These processes are complex and challenging to simplify, making sign language data acquisition and processing difficult. This challenge has significantly dampened researchers' enthusiasm and progress over a considerable period, with the majority of studies in the past decade relying on a German sign language (GSL) dataset named PHOENIX14T [25, 46] for **S**ign **L**anguage **P**roduction, **R**ecognition, and **T**ranslation tasks (SLP, SLR and SLT). Recent work [4, 86] based on the American sign language (ASL) [20, 78] and other lesser-known languages [19, 26, 32, 58, 68, 80] are relatively rare.

The existing mainstream datasets [14, 20] have significantly advanced the field. However, as time progresses, their limitations become increasingly apparent: (1) These existing datasets consist of different format files, including images, Glosses, subtitles, *etc.* These images are not easy to be directly trained. Due to redundant information in images that makes it difficult for models to learn essential pose information, training video-level SLP becomes particularly challenging. A way to reduce redundant information is to distill the gesture/posture information into text/npy/json for training. But different datasets pose extraction methods [10, 31, 51, 52, 63, 83] are different. This limitation hampers a specific format model's ability to use data from other sign languages. (2) Manual annotation for gloss is labor-intensive and time-consuming. (3) Because some videos are obtained from professionals and reprocessed into different formats, scaling up the dataset becomes exceedingly challenging. These limitations collectively impede the development and training of more advanced models.

To solve these issues, we introduce **PROMPT2SIGN**, a new multilingual dataset focusing on upper body movements in large-scale signers. The dataset overview is shown in Fig. 1 (Left), showcasing prompts, video subtitles, and files containing digital keypoints information. To create this dataset, we first process the videos using OpenPose [8] to standardize pose information in each frame. Storing keypoints information in our predefined compressed format (Compressed Pose, as shown in Fig. 1) can reduce redundancy and facilitate training with seq2seq and text2text models. Subsequently, we reduce reliance on manual annotations by auto-creating prompt words to improve cost-effectiveness. Finally, we improve the processing level of automation for the tools, making the tools highly efficient and lightweight, requiring no additional model loading to process data (*i.e.*, solving the difficulty in manual preprocessing and data collection above). Our new PROMPT2SIGN dataset is sourced from publicly available sign language datasets and videos on the Internet, covering eight sign languages, making it a comprehensive multilingual sign language dataset. More information is in Table 1 and supplementary materials.

Meanwhile, we recognize that existing models [11, 71, 73, 82, 94, 95] need improvement because training models with our new dataset brings new challenges: (1) Different sign language data cannot usually be trained simultaneously due to text-posture correspondence differences in different sign languages. (2) Handling more languages and a larger dataset results in slow and challenging training processes, with downloading, storing, and data loading difficulties. It is necessary to explore high-speed training methods. (3) The existing model structure cannot grasp more languages and understand more complex natural human conversational inputs. So, we need to explore the aspects overlooked by previous studies, such as multilingual SLP, efficient training, and the ability to understand prompts.

To overcome these challenges, we introduce **SIGNLLM**, a large multilingual Sign Language Production (SLP) model developed based on our PROMPT2SIGN dataset. It can produce the sign language representation of eight languages from texts or prompts. Our SIGNLLM has two distinct modes: (i) Multi-Language Switching Framework (MLSF), which allows multiple sign languages production in parallel by dynamically adding encoder-decoder groups. (ii) Prompt2LangGloss, allowing SIGNLLM to support static single-set encoder-decoder generation. Fig. 1 (Right) shows our model inputs and outputs, *"thank you"* is the input of mode (i), and *"how to sign 'thank you' in ASL?"* is the input of mode (ii). Two multilingual SLP modes deal with different use cases: The Multi-Language Switching Framework (MLSF) is an efficient mode without semantic confusion, like a dictionary/drawer; The Prompt2LangGloss is a user-friendly mode, like a LLM, it aims to understand complex natural language input. To address the problem of extended training time caused by more languages and a larger dataset, we utilize the Reinforcement Learning (RL) Loss concepts to quantify the quality of each training batch and prioritize valuable batches through the Priority Learning Channel.

We conduct extensive experiments and detailed ablation studies. The results validate the superior performance of our **SIGNLLM** over baseline approaches [11, 21, 43, 71, 73, 82, 94, 95] on the subsets in eight sign languages. The contributions of this paper can be summarized as follows.

- A comprehensive multilingual sign language dataset, named PROMPT2SIGN, featuring an expanded vocabulary and covering eight languages, is introduced. It is designed for broader seq2seq compatibility.
- A large multilingual Sign Language Production (SLP) model with two distinct modes—MLSF for handling text query inputs and Prompt2LangGloss for processing question-style prompts—is proposed. Our method, SIGNLLM, achieves state-of-the-art performance across eight sign languages in SLP tasks.
- We present a novel reinforcement learning-based loss function, along with a functional module named Priority Learning Channel (PLC), as a training strategy for sign language models, designed to reduce training time and computational costs.

## 2. Related Work

**Sign Language Production.** In recent years, the field of sign language research has primarily focused on Sign Language Recognition (SLR) [16, 17, 22, 29, 41, 45] and Sign Language Translation (SLT) [4, 7, 15, 44] based on deep learning. Due to a lack of suitable, high-quality datasets for ASL, deep learning researchers conducted their research [38, 71–74] based on a GSL weather theme dataset, released in 2012 [25, 46]. As previously mentioned, the data processing involved in sign language research is highly complex. Even with the release of a large-scale ASL dataset in 2021 [20], work focused on ASL-related themes based on it has not emerged quickly, as existing work is not easily transferable. The situation is worse for minority languages.

**Large Language Models.** LLMs refer to giant transformer models trained on extensive textual data, exhibit capabilities in understanding natural language and addressing complex tasks [5, 13, 77, 87, 89]. Sign language is a visual language, theoretically different from language models. However, most current work uses text2text and seq2seq models [48, 61, 66, 99], converting key points/dense maps/grid poses into sequences for training, as opposed to directly training images. Hence, viewing the core process of SLP, text2pose, as a language model is justifiable. Extensive research indicates that an increase in parameters or data volume [35, 42] significantly enhances the abilities of LLMs [5, 13, 65]. There are more than a hundred sign languages in the world, most of which have datasets in video form. Therefore, conducting advanced research to address the anticipated surge in data volume in the future is of paramount importance. In this work, we aim to enable the model to generate sign languages across diverse linguistic backgrounds, ensuring its adaptability to our new dataset.

## 3. Our Benchmark: Prompt2Sign

**Existing Datasets Weaknesses.** The main shortcoming of previous work is the lack of unified data storage formats, when there is a mismatch between these models in Sign Language Production (SLP) [11, 71, 73, 82, 94, 95] and Sign Language Translation (SLT) [4, 7, 15, 44], it can lead to complex challenges: (1) The results of the SLT model are difficult to use as training data for the SLP model directly due to format incompatibility (*e.g.*, [6, 86] & [70, 75]) (2) The results of the SLP model are difficult to use as input for the SLT model (evaluation experiment needs, *e.g.*, [6] & [4]). (3) The output of the SLP model is not suitable as input for most style transfer models (*e.g.*, [11, 97, 107], the researchers have to train a pose2video model themselves). Therefore, we create a standardized dataset to address data collection, utilization, and storage challenges.

**Data Collection.** Our data collection process, in contrast to previous methods, includes the following steps: (1) downloading sign language videos in specific languages from the Internet and public datasets [19, 26, 32, 58, 68, 80]; (2) editing and aligning these videos; (3) extracting 2D keypoints from each video frame using OpenPose [8] and saving them as JSON files; (4) calculating the 3D poses, which are stored in a predefined compressed data format. See further details below and in the *supplementary materials*.

**Dataset Range.** We choose How2Sign [19], PHOENIX-14T [26], KSL [100], Signsuisse [58] (contains 3 languages), LSA64 [68], AuTSL [80], and some of the more popular works were not considered due to their limited accessibility and potential usage restrictions. Additionally, some available multilingual datasets [30, 34, 55, 101] may not possess the same level of comprehensiveness as ours. For example, [30] and [101] translate two types of sign language videos into spoken language (SLT), while our work is from spoken language to videos (SLP). We aim to establish a robust multilingual SLP method with a dataset that supports a broader range of application scenarios. For example, [30] is limited to Bible translation, and [101] focuses solely on cross-lingual SLT. In contrast, our work covers a broader range of scenarios and comprehensively addresses the SLP task. Additionally, some datasets, such as [55] and [34], are restricted to multilingual dictionary forms.

| Subset | ASL | GSL | DSGS | LSF-CH | LIS-CH | LSA | KSL | TSL |
|---|---|---|---|---|---|---|---|---|
| Train | 31,047 | 7,096 | 8,043 | 5,672 | 2,254 | 2,400 | 700 | 28,142 |
| Dev | 1,739 | 519 | 500 | 500 | 250 | 400 | 300 | 4,418 |
| Test | 2,343 | 642 | 500 | 250 | 250 | 400 | 200 | 3,742 |

Table 1. **Dataset Statistics** [19, 26, 57, 58, 68, 80, 100] **:** The number of video clips in their train, dev, test set, respectively. Languages included: American (ASL), German (GSL, Alias DGS), Swiss German (DSGS), French Sign Language of Switzerland (LSF-CH), Italian Sign Language of Switzerland (LIS-CH), Argentine (Lengua de Señas Argentina, LSA), Korean (KSL), and Turkish (TSL). "Sign Language" is omitted in some full names.

**Unified Compression of Pose Data.** Building on previous work [71, 75, 104], we develop a three-step tool for standardizing data processing. The tool is highly efficient and lightweight, requiring no additional model loading and supporting large-scale data processing. Furthermore, we optimize it specifically for sign language data processing (e.g., removing unnecessary leg movement computations and integrating these optimizations into PROMPT2SIGN's pipeline). The main steps of Data Collection can be visualized as: 2D Keypoints to 3D Keypoints to Compressed Pose. Among all the steps, the most crucial part is the transition from 2D to 3D compressed pose data:

`Step I`: First, we obtain the length of the skeleton through the 2D keypoint coordinates ($x$ and $y$), $a$ and $b$ represent indices that identify the two keypoints (or joints) forming a bone $L = \sqrt{(ax - bx)^2 + (ay - by)^2}$.

`Step II`: We compute the 3D rotation angles from 2D keypoints data: $A_x, A_y, A_z = \frac{\text{angle}_x, \text{angle}_y, \text{angle}_z}{\sqrt{\text{angle}_x^2 + \text{angle}_y^2 + \text{angle}_z^2}}$, where $A$ represents the normalized angles.

`Step III`: We define $P_x, P_y, P_z$ as the starting joint coordinates computed from 2D keypoints, and define $Q_x, Q_y, Q_z$ as the target coordinates in 3D space: $Q_x = P_x + L \times A_x, Q_y = P_y + L \times A_y, Q_z = P_z + L \times A_z$.

These mathematical formulas initialize the skeletal model by calculating skeletal length $L$, root node position, rotation angle, and 3D coordinates $(x, y, z)$, serving as input for simulating 3D human skeletal motion. Compared to previous methods [104], this approach is much more efficient, focusing solely on extracting the posture and gesture information relevant to sign language from the video data. This preprocessing step discards redundant information, reduces data size by 80% relative to the raw video, and standardizes the format, enabling easier integration with text-to-text and sequence-to-sequence models and applications without the need for sign language-specific data loaders.

**Dataset Statistics.** After the data processing, train, dev, and test sets of different language parts are shown in Table 1. We constructed 120 English templates and 210

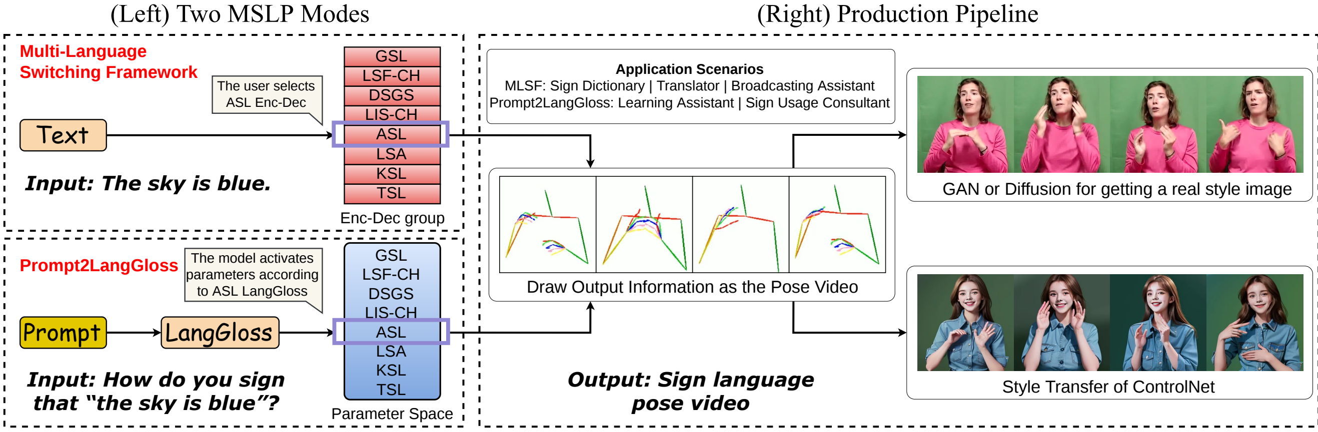


Figure 2. (Left) MLSF contains parallel Enc-Dec groups (*i.e.*, *Text2Pose* $\times$ *number of languages*), the Prompt2LangGloss adds a language attribute marker at the gloss channel (*i.e.*, *Text2Gloss2Pose* $\rightarrow$ *Prompt2LangGloss2Pose*). (Right) The output of SIGNLLM can be converted into a skeletal pose video, which can then be rendered into a realistic human appearance by vid2vid models [11, 69, 97, 105, 107].

prompt word templates for other languages (with 30 templates for each language), which are randomly associated with the oral text data to form the prompt part. The templates are carefully selected from hundreds of sentences generated by LLM that cover most everyday expressions.

# 4. Our Model: SignLLM

## 4.1. Preliminary of Text2Pose Method

The general SLP pipeline (*i.e.*, text to sign language video) [38, 71–74] has following steps: text-to-gloss conversion, gloss-to-pose mapping, and finally pose-to-video rendering. In our work, we mainly focus on the first two steps.

**Text2Gloss & Gloss2Pose.** Essentially, the transformation from text-to-gloss and gloss-to-pose can be distilled to a sequence-to-sequence [71, 72] problem in the realm of textual data, and their structures to bear significant resemblances. We define $x_u$ as the input text $x$ tokens at position $u$ (total number is $\mathcal{U}$, position from 1 to $\mathcal{U}$), $p_w$ as the output pose $p$ at position $w$ (total number is $\mathcal{W}$, frame position from 1 to $W$), and then we use an encoder-decoder transformer framework to convert input into output:

$$f_u = Enc_{input2output}(x_u | x_{1:\mathcal{U}}) \tag{1}$$

$$p_{w+1} = Dec_{input2output}(p_w | p_{1:w-1}, f_{1:\mathcal{U}}) \tag{2}$$

Here, $f_u$ denotes the encoded source of $x_u$. The output text tokens generated from this process form the input for the next stage of our translation model. In short, text2pose is the core method of SLP, and some researchers use gloss as an intermediate to make it text2gloss2pose.

## 4.2. Design Overview

**Architecture.** SIGNLLM has two modes Multi-Language Switching Framework (MLSF) and Prompt2LangGloss, as shown in Fig. 2 (Left), both make the model capable of multilingual sign language production by using the multilingual PROMPT2SIGN dataset. Both modes can be trained by new RL Loss and Priority Learning Channel module.

**Motivation.** Unlike existing SLP models, LLM directly fine-tuning for text-to-pose translation would hamper its dialogue capabilities, while using it without fine-tuning would treat translation requests as questions rather than performing actual text-to-pose translations. Therefore, we propose two specialized modes: MLSF for direct text-to-pose translation and Prompt2LangGloss for LLM-based interaction (*e.g.*, *"how to sign 'thank you' in ASL?" as input*). These modes, analogous to evolutionary branches of the same species, serve complementary purposes in multilingual SLP.

## 4.3. Two Multilingual SLP Modes

MLSF is a mode that most existing models can refer to, like a dictionary/translator, and Prompt2LangGloss mode is designed specifically for sign language LLM. They are like two evolutionary branches diverged from the same species.

**Multi-Language Switching Framework.** It can be understood as having multiple parallel Text2Pose channels/Enc-Dec groups, each language has an Enc-Dec group, allowing each channel/Enc-Dec to be independently trained and inferred. Text2Pose visual representation is shown on the left of Fig. 2, the red rectangle represents the eight Enc-Dec in our model, and the middle partition represents the parameters stored separately in different Enc-Dec groups. The assignment operation could be formalized as Enc$_{\mathcal{L}} = \mathcal{E}_{\mathcal{L}}$ and Dec$_{\mathcal{L}} = \mathcal{D}_{\mathcal{L}}$. Here, $\mathcal{L}$ denotes the language of input, while $\mathcal{E}_{\mathcal{L}}$ and $\mathcal{D}_{\mathcal{L}}$ are the mapping from language $\mathcal{L}$ to an encoder and decoder in the sets $\mathcal{E}$ and $\mathcal{D}$. Similar to selecting tools from a drawer, MLSF allows you to choose the appropriate $\mathcal{E}_{ASL}$-$\mathcal{D}_{ASL}$ pair from the $\mathcal{E}$-$\mathcal{D}$ groups for training and inference. This modular design functions like a language drawer system, where each language Enc-Dec component can be accessed and utilized on demand from eight pairs.

**Prompt2LangGloss.** While MLSF tackles multilingual support through architectural design, Prompt2LangGloss approaches the challenge from a linguistic perspective. This mode introduces a novel intermediate representation that bridges the gap between natural language understanding and sign language generation. Enriching the traditional gloss notation with language-specific attributes creates a more nuanced and contextually aware translation process.

Gloss, essentially a shorter textual representation of sign language gestures, operates as an intermediate entity when using a text2pose model. As shown in Fig. 2 (Left), our proposed enhancement of this model involves appending an additional language attribute to each text word during the reading and tokenizing stages. For instance, a traditional gloss token “<xxx>” can be transformed into “<ASL_xxx>”, thus introducing a LangGloss layer of conditional input $f_u = Enc_{t2lg}(x_u|x_{1:\mathcal{U}})$ into SLP based on Eq. (1): $lg_{w+1} = Dec_{t2lg}(lg_w|lg_{1:w-1}, f_{1:\mathcal{U}})$. So, our LangGloss is a pseudo-Gloss used to distinguish language information in the parameter space by identifying language attributes. This way, we solve several challenges: (1) LangGloss allows the existing models to train multilingual data. By adding language attributes to gloss, we reduce the semantic ambiguity that occurs when words share the same form across different languages. (2) LangGloss, as a mediator, can solve the limitations of the existing model in understanding complex, natural human inputs. It reduces the negative impact of directly processing intricate prompt words, improving the model's response accuracy.

In short, these two modes create a robust multilingual SLP system: This dual-mode design enables SIGNLLM to achieve processing efficiency and semantic accuracy in existing models and sign language LLMs, respectively.

### 4.4. Reinforcement Learning Training Strategy

To reduce training time, we utilize the RL reward concept to quantify each training batch's quality and prioritize valuable batches through the Priority Learning Channel module. Both Multilingual SLP modes can use RL Loss and PLC.

**Reinforcement Learning Loss.** Reinforcement Learning (RL) has an important advantage in identifying high-value actions or samples, allowing for prioritized learning of valuable data. This approach could help address the challenge of slow training when using more sign languages. But before learning high-value data batch, we should transform the ordinary generative model first into an RL-like model, so we design RL Loss for model transformation.

Concretely, we set the input sequence as the state $s_t$, and the output sequence is the action $a_t$, and the reward $r_t$. $t$ stands for time, $i$ stands sample, $\hat{y}_i$ represents the model's predicted output, and $N$ denotes the total number of samples. The closer the prediction is to reality (mean squared error), the greater the reward: $r = -\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$.

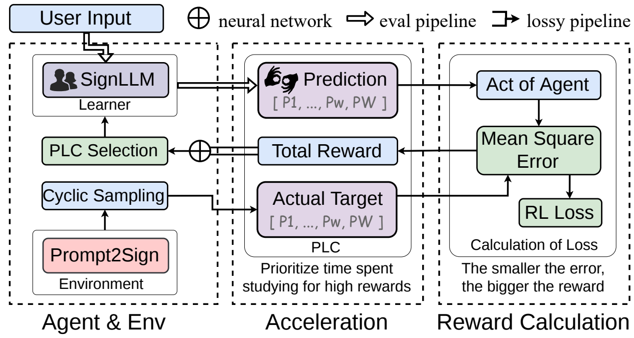


Figure 3. RL elements: *User, Agent, Environment, Cyclic Sampling, PLC* to sketch the sequence prediction learning process.

With this interpretation, we can reformulate the traditional supervised learning problem of minimizing MSE loss to maximize the expected cumulative reward, where $L$ denotes the MSE loss function, $M$ is the model, and $x_t$, $y_t$ are the model inputs and corresponding targets respectively:

$$\theta^* = \underset{\theta}{\text{argmax}}\, E_\theta\left[\sum_{t=0}^{T} r_t\right] = \underset{\theta}{\text{argmin}}\, E_\theta\left[\sum_{t=0}^{T} L(y_t, M(x_t))\right] \quad (3)$$

where $\theta$ represents the trainable parameters of the model, $E_\theta$ denotes the expected value with respect to the model parameters $\theta$, and $T$ is the total number of time steps in the sequence. Here, argmax/argmin returns the value of $\theta$ that maximizes/minimizes the expected cumulative reward. These optimized parameters $\theta$ are founded by using gradient descent, updating parameters proportionally to the gradient of expected cumulative reward concerning model parameters. In summary, we quantify the effectiveness of MSE optimization at each time step as reward $r$. Then the $r$ will serve as the input for our Priority Learning Channel.

**Priority Learning Channel.** The RL Loss itself does not possess subjective acceleration capabilities, it is designed for PLC to prioritize the learning of more valuable data. We have defined rewards $r$, sample $i$, and data for each batch $j$. They then are converted into sampling probabilities for each data sample according to $P(i) = \frac{r(i)^\eta}{\sum_{j\in S} r(j)^\eta}$, where $\eta$ regulates the intensity of prioritization, and $S$ represents the dataset. By employing these sampling probabilities, the choice of data samples for each batch is no longer uniform but regulated by their respective rewards (*e.g.*, if the reward is less than 50%, skip the batch). The per-step RL loss, $L(i)$, is computed for the chosen data, which is then used to optimize the model parameters following the policy gradient theorem. This procedure is formally expressed as $Minimize\ \ E_{i\sim P(i)}[L(i)]$ (where $E$ is the expectation), the whole RL Loss and PLC system is shown in Fig. 3. By continually updating the model based on the most rewarding samples, the PLC brings the advantages of RL to sequence prediction tasks. The adaptive nature of the PLC ensures that the model's focus shifts by the model's evolving knowledge, thereby accelerating the learning process.

| Type: | DEV SET BLEU-4 | DEV SET BLEU-3 | DEV SET BLEU-2 | DEV SET BLEU-1 | DEV SET ROUGE | TEST SET BLEU-4 | TEST SET BLEU-3 | TEST SET BLEU-2 | TEST SET BLEU-1 | TEST SET ROUGE |
|---|---|---|---|---|---|---|---|---|---|---|
| NSLP-G [39] | - | - | - | - | - | 5.75 | 8.21 | 11.62 | 17.55 | 31.98 |
| Fast-SLP Transformers [? ] | 17.19 | 23.11 | 29.49 | 36.96 | 55.85 | 12.85 | 17.35 | 23.38 | 39.46 | 46.89 |
| Neural Sign Actors [2] | - | - | - | - | - | 13.12 | 18.25 | 25.44 | 41.31 | 47.55 |
| **SignLLM-1x40M-Base-M (ASL)** | 18.77 | 25.42 | 32.44 | 40.66 | 61.44 | 14.13 | 19.08 | 25.72 | 43.40 | 51.57 |
| **SignLLM-1x120M-Large-M (ASL)** | 19.40 | 26.11 | 33.38 | 41.89 | 63.21 | 14.52 | 19.65 | 26.44 | 44.72 | 53.14 |
| **SignLLM-1x1B-Super-M (ASL)** | **20.09** +2.90 | **27.04** +3.93 | **34.45** +4.96 | **43.16** +6.20 | **65.19** +9.36 | **15.03** +2.18 | **20.28** +2.93 | **27.35** +3.97 | **46.17** +6.71 | **54.86** +7.97 |
| **SignLLM-1x40M-Base-P (ASL)** | 17.34 | 23.57 | 29.87 | 37.81 | 56.93 | 13.06 | 17.66 | 23.77 | 40.15 | 47.76 |
| **SignLLM-1x120M-Large-P (ASL)** | 18.05 | 24.28 | 31.04 | 38.97 | 58.78 | 13.48 | 18.27 | 24.57 | 41.57 | 49.42 |
| **SignLLM-1x1B-Super-P (ASL)** | 18.68 +1.49 | 25.11 +2.00 | 31.99 +2.50 | 40.14 +7.18 | 60.47 +4.62 | 13.93 +0.92 | 18.86 +1.51 | 25.40 +2.02 | 42.87 +3.41 | 50.91 +4.02 |

Table 2. **American Sign Language Production (ASLP):** Comparison of SignLLM variants with baseline on *Text to Pose* task by using our Prompt2Sign ASL part. "-" : The NSA [2] and NSLP-G [39] have not been tested on the dev set, and there is no source code. The improvement (**+num**) is relative to the latest work [**?** ].

| Type: | Language: | DEV SET BLEU-4 | DEV SET BLEU-3 | DEV SET BLEU-2 | DEV SET BLEU-1 | DEV SET ROUGE | TEST SET BLEU-4 | TEST SET BLEU-3 | TEST SET BLEU-2 | TEST SET BLEU-1 | TEST SET ROUGE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **SignLLM-6x40M-Base-M** | DSGS | 9.73 | 15.82 | 19.85 | 24.84 | 37.57 | 7.34 | 9.89 | 16.85 | 26.04 | 31.51 |
| **SignLLM-6x120M-Large-M** | DSGS | **11.45** | 17.28 | 19.84 | 29.07 | **41.69** | **9.69** | 14.06 | 16.35 | 29.80 | **31.51** |
| **SignLLM-6x40M-Base-M** | LSF-CH | 9.79 | 18.48 | 23.13 | 28.86 | 34.98 | 8.92 | 13.12 | 16.11 | 26.48 | 37.92 |
| **SignLLM-6x120M-Large-M** | LSF-CH | **13.72** | 20.79 | 23.40 | 25.15 | **38.39** | **9.60** | 12.58 | 16.98 | 22.71 | **41.96** |
| **SignLLM-6x40M-Base-M** | LIS-CH | 10.81 | 14.46 | 19.93 | 24.55 | 35.83 | 7.34 | 10.56 | 15.24 | 22.73 | 36.42 |
| **SignLLM-6x120M-Large-M** | LIS-CH | **12.10** | 18.04 | 23.01 | 25.95 | **36.98** | **9.30** | 11.20 | 15.68 | 23.38 | **38.37** |
| **SignLLM-6x40M-Base-M** | LSA | 10.72 | 15.55 | 21.76 | 25.91 | 38.78 | 7.33 | 14.86 | 16.68 | 22.55 | 34.42 |
| **SignLLM-6x120M-Large-M** | LSA | **11.69** | 14.79 | 26.25 | 28.08 | **39.01** | **8.21** | 11.04 | 17.05 | 26.68 | **37.46** |
| **SignLLM-6x40M-Base-M** | KSL | 9.42 | 14.67 | 17.24 | 26.41 | 31.96 | 8.31 | 11.84 | 17.93 | 24.15 | **33.78** |
| **SignLLM-6x120M-Large-M** | KSL | **12.91** | 19.45 | 15.17 | 24.07 | **37.83** | **10.09** | 13.06 | 18.37 | 25.75 | 33.69 |
| Hybrid Translation System [43] | TSL | - | - | - | - | - | 12.64 | 18.28 | 31.48 | 53.17 | - |
| **SignLLM-6x40M-Base-M** | TSL | 14.53 | 19.86 | 29.93 | 36.86 | 58.01 | 13.23 | 17.80 | 25.39 | 39.30 | 57.03 |
| **SignLLM-6x120M-Large-M** | TSL | **15.17** | 21.70 | 31.73 | 38.86 | **71.10** | **14.36** | 18.74 | 26.96 | 43.21 | **57.12** |

Table 3. **Multilingual Sign Language Production (MSLP):** Comparison of different SignLLM M-mode variants with a baseline on *Text to Pose* task. We propose the first Multilingual SLP benchmark, with the exception of the existing TSL-Baseline.

## 5. Experiments and Discussions

**Setup.** We provide the naming rules as follows: SignLLM-{number of languages}x{single language parameters}-{submode size}-{the mode of training}-{the language of input}. Such as "SignLLM-2x40M-Base-M (ASL)", the nomenclature "2x40" denotes that the model comprises 2 language knowledge, with each language component estimated to be around 40 million parameters in size, and a total is 80 million parameters ("1B" represents a total of 1 billion parameters). There are Base, Large, and Super versions, depending on a single language parameter size provided by the model. The encoder and decoder of our model versions (*i.e.*, Base, Large, Super) both have two layers. When the model is expanded to Large and Super versions, the layers are unchanged, the parameters are expanded by about two and four times, respectively. M and P stand for models trained using MLSF and Prompt2LangGloss. At the end is the language of the input model, ASL, GSL, LSA *etc*.

**Metrics.** `(i) BLEU-n score` measures the similarity between machine-generated translations and reference translations based on n-grams, the closer the predicted result is to the input (reference), the higher the value. BLEU-n [60]

| Approach: | DEV SET BLEU-4 | DEV SET ROUGE | TEST SET BLEU-4 | TEST SET ROUGE |
|---|---|---|---|---|
| Progressive Transformers [71] | 11.82 | 33.18 | 10.51 | 32.46 |
| Adversarial Training [70] | 12.65 | 33.68 | 10.81 | 32.74 |
| Mixture Density Networks [72] | 11.54 | 33.40 | 11.68 | 33.19 |
| Mixture of Motion Primitives [73] | 14.03 | 37.76 | 13.30 | 36.77 |
| Photo-realistic SLP [75] | 16.92 | 35.74 | 21.10 | 42.57 |
| Fast-SLP Transformers [? ] | 18.26 | 39.62 | 22.15 | 46.82 |
| **SignLLM-1x40M-Base-M (GSL)** | 18.61 | 40.69 | 22.76 | 48.05 |
| **SignLLM-1x120M-Large-M (GSL)** | **19.31** +1.05 | 41.42 | **23.25** +1.10 | 49.08 |
| **SignLLM-1x1B-Super-M (GSL)** | 19.07 | **41.83** +2.21 | 23.21 | **49.52** +2.70 |
| **SignLLM-1x40M-Base-P (GSL)** | 17.12 | 37.43 | 20.93 | 44.21 |
| **SignLLM-1x120M-Large-P (GSL)** | 17.55 | 38.10 | 21.39 | 45.16 |
| **SignLLM-1x1B-Super-P (GSL)** | 17.54 | 38.48 | 21.35 | 45.57 |

Table 4. **German Sign Language Production (SLP):** Comparison of different models with existing work on *Text to Pose* task. The improvement (**+num**) is relative to the latest work [**?** ].

means that n words are used as the basic computing unit, and the higher the n, the higher the fluency requirement. `(ii) ROUGE score` [49] is similar to BLEU, but it is more concerned with consistency and coverage. It indicates better agreement between the generated and reference texts, indicating a more accurate and comprehensive summary. `(iii) DTW score` underpinned by dynamic programming principles [3], is employed to ascertain the smallest manipulation distance between clips and sentences; the lower, the better.

| Approach: | DEV SET BLEU-4 | BLEU-3 | BLEU-2 | BLEU-1 | ROUGE | TEST SET BLEU-4 | BLEU-3 | BLEU-2 | BLEU-1 | ROUGE |
|---|---|---|---|---|---|---|---|---|---|---|
| **Base + Normal MSE Loss** | 10.96 | 14.68 | 22.49 | 40.29 | 44.13 | 9.27 | 10.72 | 18.64 | 39.88 | 40.39 |
| **Base + RL Loss** | 18.33 | 24.28 | 32.18 | 48.83 | 52.61 | 13.26 | 17.33 | 24.51 | 40.76 | 40.95 |
| **Base + RL Loss & PLC** | **18.77** | **25.42** | **32.44** | **52.66** | **61.44** | 14.13 | 19.08 | 25.72 | 43.40 | **51.57** |
| **Base + MSE + Prompt2LangGloss** | 15.78 | 22.66 | 30.27 | 48.30 | 50.88 | 16.32 | 20.27 | 28.70 | 47.89 | 48.18 |
| **Base + MSE + MLSF** | 16.44 | 23.79 | 32.32 | 50.97 | 53.44 | **17.17** | **21.25** | **30.12** | **50.25** | 50.47 |

Table 5. **Comparison of Different Modules:** Base: SIGNLLM-40M-Base results for *Text to Pose* task on the ASL part of PROMPT2SIGN. PLC: Priority Learning Channel. Top performances are highlighted in **bold**, while second top performances are underlined.

### 5.1. Quantitative Evaluation

**Back Translation.** Back translation means translating generated sign language videos back into spoken language sentences. These sentences are compared with the original input sentences to evaluate the translation quality. The task is widely adopted to evaluate Sign Language Production (SLP) as it can indicate the accuracy of the produced sign language videos [71], our translation models [15] are trained on the corresponding language Prompt2Sign pose data, the performance is between bleu-4 22.3 and 27.6, it is comparable to the translation model used in previous work.

In Table 2, we conduct American Sign Language Production back-translation tests using SIGNLLM on the ASL part of our new dataset, and Table 4 further compares our method with other recent approaches for German SLP on the GSL part of PROMPT2SIGN dataset [23, 70–73]. These two languages stand for high-resource languages (*i.e.*, languages with rich data resources), and our comprehensive tests at different levels on the dataset demonstrate impressive performance compared with the latest works in the field. The results affirm the competitiveness and potential of our proposed method, regardless of the specific sign language in use. More evaluation results and analyses for these two languages can be found in the supplementary materials.

In Table 3, we present the results of our model for six different sign languages. These six languages represent low-resource languages, and they are considered limited languages. For languages with low resources, their vocabulary, video time, and diverse corpus sources are relatively low, making training more difficult. From the table data, it can be observed that our performance remains strong in languages where training data is lacking. As long as the input text/prompt can be encoded as a computationally recognizable word and video exists, our method can translate it into the corresponding language pose video after training.

### 5.2. Ablation Evaluation

**Performance.** Ablation results in Table 5 indicates our four innovative strategies (Prompt2LangGloss, MLSF modes, RL Loss and PLC module) significantly improve the model's performance, and four strategies contribute to substantial improvements. When replacing the standard MSE Loss with RL Loss, we observe a significant improvement in the model's performance. These results show that PLC has successfully enhanced the adaptability of the original RL Loss to unknown environments. Incorporating the PLC module further enhances these gains. We also note that the base model with MSE, when upgraded with Prompt2LangGloss, achieves higher scores with the new mode, slightly underperforming compared to the MLSF mode but far surpassing the baseline model in the first row. This indicates that, with sufficient data, the Prompt2LangGloss mode demonstrates good usability, and we explore this further in the supplementary materials.

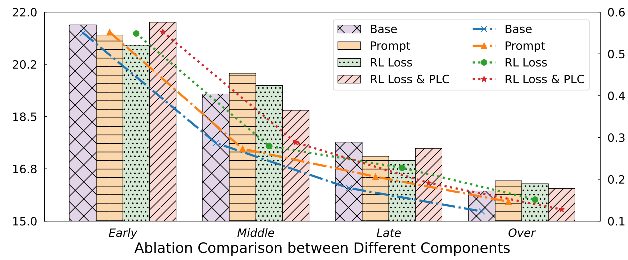


Figure 4. **RL Training Efficiency Analysis:** Comparison of different Settings on DTW values (the lower the better) at different peroid (every 30 epochs a peroid). Left Y-axis: Value of DTW. Right Y-axis: Value of Loss. Prompt: Prompt2LangGloss mode.

In Fig. 4, we compare the Base model's performance with different module schemes. The ablation comparison is primarily based on observing the variability of DTW scores across epochs to assess the effectiveness of each approach. We observe that: (i) The standalone use of the Prompt2LangGloss model exhibits the lowest efficiency, as it introduces noise by incorporating prompts and tokenizers, which is also evident in Table 5. (ii) The combination of the two RL methods shows modest performance improvement compared to using each method individually, although it is not statistically significant. Future work could explore adaptive weighting mechanisms or hierarchical integration of these methods to leverage their complementary strengths fully. (iii) In terms of training efficiency, they significantly reduce training time. Compared to the base setting, our RL Loss with PLC approach reduces training time by 27.1%, which is crucial for training large-scale sign language models. Our ablation studies reveal that RL-based sample prioritization not only enhances training efficiency but also provides a scalable, unified, multilingual SLP framework.


Figure 5. We use an adjusted vid2vid model [24] to convert the predicted skeletal pose video into a more realistic final video.

### 5.3. Qualitative Evaluation

**Qualitative Presentation.** We use our predicted pose results as input, which are then used to generate rendered videos in Fig. 5. We can observe that our video outcomes are of high quality, with highly accurate finger movements and high image fidelity. Our results surpass all previous works, benefiting not only from technological advancements but also from the superior output quality of our SIGNLLM compared to previous smaller models: the postures we predict are rarely missing, unlike previous works that often suffered from issues such as flickering, incomplete or missing fingers, and low input quality due to densely packed fingers, so it works well with the latest model [24]. However, the final sign language video finger problem still exists anyway, and in the future, we can consider the special optimization of these style transfer vid2vid models.

| Approach: | DEV SET BLEU-4 | DEV SET ROUGE | TEST SET BLEU-4 | TEST SET ROUGE |
|---|---|---|---|---|
| Progressive Transformers [71] | 10.79 | 36.15 | 9.59 | 35.42 |
| Fast-SLP Transformers [? ] | 16.68 | 43.2 | 24.24 | 51.05 |
| **SignLLM-1x40M-Base-M (GSL)** | 16.96 | 44.41 | 24.74 | 52.45 |
| **SignLLM-1x120M-Large-M (GSL)** | **17.73** +1.05 | 45.11 | **25.39** +1.15 | 53.45 |
| **SignLLM-1x40M-Base-P (GSL)** | 16.27 | 43.76 | 24.12 | **54.35** +3.30 |
| **SignLLM-1x120M-Large-P (GSL)** | 16.71 | **45.40** +2.20 | 25.04 | 52.80 |

Table 6. Presentation Effect Study: Results of *Text to Sign* task in GSL. The improvement (**+num**) is relative to the latest work [**?** ].

In Table 6, we conducted a series of final video back-translation evaluations (as shown in Fig. 5) based on the German SLP task to investigate two main research questions: (i) Whether rendering the predicted results into real sign language videos would lead to a decrease in accuracy. (ii) How does our work compare to previous studies on the task of text-to-sign-language real video generation? Based on our observations, there was generally not a significant accuracy loss, but there were some fluctuations compared to the baseline. Our approach outperformed previous works, which could be attributed to the higher quality of our data, making it more suitable as input for style transfer models.

### 5.4. Discussion

**Societal Impact.** Our model has the potential to assist people with disabilities in three key areas: sign language teaching, generative sign language translation, and real-time interpretation for broadcasting. (1) Traditional sign language teaching relies heavily on human instructors and static pictorial representations, limiting learning accessibility. (2) Current sign language translation software remains inadequate for effective communication between deaf people and their family members who lack sign language knowledge. (3) Additionally, real-time sign language interpretation is only available for limited content like major news broadcasts, creating barriers for the deaf community in daily life. However, its current level of accuracy is not high enough to be fully trusted, and users must be cautious to use it.

**Limitation.** Our tool enhances sign language data processing automation but isn't fully end-to-end. Manual preprocessing is still required for OpenPose processing, video editing, and transcript alignment, *etc*. For pose video to final video conversion, when we use a style transfer model to make the final video, it requires more complex processing of the pose video. So there is still a way to go before large-scale use, which requires us to make industrial-level adaptation improvements to the output style of SIGNLLM.

## 6. Conclusion

We present SIGNLLM, a large multilingual SLP model. For training this model, we propose PROMPT2SIGN, a standardized dataset that contains eight sign languages. Our model with two modes, MLSF and Prompt2LangGloss, progressively incorporates more sign languages while preserving translation efficiency and introducing LLM capabilities. Our new RL loss and new PLC module solve the challenge of longer training time due to more data. Finally, we show baseline comparisons, ablation studies, experiments under various parameters, and qualitative evaluations for discussion, which proves the efficacy of our methodology.

# References


[1] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing Images of Humans in Unseen Poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[2] 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 (CVPR)*, 2024. 6

[3] Donald J Berndt and James Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. In *AAAI-94 Workshop on Knowledge Discovery in Databases*, 1994. 6

[4] Matyáš Boháček and Marek Hrúz. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 182–191, 2022. 1, 2, 3

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 3

[6] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Smpler-x: Scaling up expressive human pose and shape estimation, 2023. 1, 3, 5

[7] Necati Cihan Camgöz, 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 (CVPR)*, 2018. 2, 3

[8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3

[9] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3

[10] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[11] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody Dance Now. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2019. 2, 3, 4, 1

[12] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 1, 5

[13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022. 2, 3

[14] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, pages 7784–7793, 2018. 1, 3

[15] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 7

[16] Helen Cooper and Richard Bowden. Large Lexicon Detection of Sign Language. In *International Workshop on Human-Computer Interaction*, 2007. 2

[17] Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[18] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[19] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 3

[20] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Tor-

res, and Xavier Giro-i Nieto. How2Sign: A Large-Scale Multimodal Dataset for Continuous American Sign Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3

[21] Sarah Ebling. *Automatic Translation from German to Synthesized Swiss German Sign Language*. PhD thesis, University of Zurich, 2016. 2

[22] Sen Fang, Chunyu Sui, Hongwei Yi, Carol Neidle, and Dimitris N. Metaxas. Signx: The foundation model for sign recognition, 2025. 2

[23] Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Yapeng Tian, and Chen Chen. Signdiff: Diffusion model for american sign language production, 2025. 7

[24] Mengyang Feng, Jinlin Liu, Kai Yu, Yuan Yao, Zheng Hui, Xiefan Guo, Xianhui Lin, Haolan Xue, Chen Shi, Xiaowen Li, Aojie Li, Xiaoyang Kang, Biwen Lei, Miaomiao Cui, Peiran Ren, and Xuansong Xie. Dreamoving: A human video generation framework based on diffusion models, 2023. 8

[25] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2012. 1, 2

[26] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, 2014. 1, 3

[27] Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. Llms are good sign language translators, 2024. 7

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2014. 1

[29] Kirsti Grobel and Marcell Assan. Isolated Sign Language Recognition using Hidden Markov Models. In *IEEE International Conference on Systems, Man, and Cybernetics*, 1997. 2

[30] Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. Jwsign: A highly multilingual corpus of bible translations for more diversity in sign language processing. *arXiv preprint arXiv:2311.10174*, 2023. 3, 6

[31] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 2

[32] Soomin Ham, Kibaek Park, YeongJun Jang, Youngtaek Oh, Seokmin Yun, Sukwon Yoon, Chango Jo Kim, Han-Mu Park, and In So Kweon. Ksl-guide: A large-scale korean sign language dataset including interrogative sentences for guiding the deaf and hard-of-hearing. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2021. 1, 3

[33] Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. Extending the public dgs corpus in size and depth. In *LREC2020 - Workshop on the Representation and Processing of Sign Languages*, pages 75–82, 2020. 3

[34] Marlene Hilzensauer and Klaudia Krammer. A multilingual dictionary for sign languages:" spreadthesign". In *ICERI2015 Proceedings*, pages 7826–7834. IATED, 2015. 3, 6

[35] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. abs/2203.15556, 2022. 3, 5

[36] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[37] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *AAAI*, 2018. 3

[38] 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*, 2021. 1, 2, 4

[39] Eui Jun Hwang, Jung-Ho Kim, and Jong C. Park. Non-autoregressive sign language production with gaussian space. In *The 32nd British Machine Vision Conference (BMVC 21)*. British Machine Vision Conference (BMVC), 2021. 6

[40] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 5

[41] Timor Kadir, Richard Bowden, Eng-Jon Ong, and Andrew Zisserman. Minimal Training, Large Lexicon, Unconstrained Sign Language Recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2004. 2

[42] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. 3, 5

[43] Dilek Kayahan and Tunga Güngör. A hybrid translation system from turkish spoken language to turkish sign language. In *2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6, 2019. 2, 6

[44] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural Sign Language Translation based on Human Keypoint Estimation. *Applied Sciences*, 2019. 2, 3

[45] Oscar Koller. Quantitative Survey of the State of the Art in Sign Language Recognition. *arXiv preprint arXiv:2008.09918*, 2020. 2

[46] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015. 1, 2

[47] Marek Kowalski, Stephan J Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. CONFIG: Controllable Neural Face Image Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[48] Julia Kreutzer, Joost Bastings, and Stefan Riezler. Joey NMT: A Minimalist NMT Toolkit for Novices. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 3

[49] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics, 2004. 6

[50] Yahui Liu, Marco De Nadai, Gloria Zen, Nicu Sebe, and Bruno Lepri. Gesture-to-Gesture Translation in the Wild via Category-Independent Conditional Maps. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2019. 1

[51] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM TOG*, 2015. 2, 4

[52] 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. 1, 2

[53] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose Guided Person Image Generation. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 1

[54] Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-Consistent Video-to-Video Synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 5

[55] Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert, and Eva Safar. Dicta-sign-building a multilingual sign language corpus. In *5th Workshop on the Representation and Processing of Sign languages: Interactions between Corpus and Lexicon. Satellite Workshop to the eighth International Conference on language Resources and Evaluation (LREC-2012)*, 2012. 3, 6

[56] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable Person Image Synthesis with Attribute-Decomposed GAN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[57] Amit Moryossef and Mathias Müller. Sign language datasets. `https://github.com/sign-language-processing/datasets`, 2021. 3

[58] Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore, 2023. Association for Computational Linguistics. 1, 3

[59] Carol Neidle and Christian Vogler. A new web interface to facilitate access to corpora: Development of the asllrp data access interface (dai). In *Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC*, 2012. 3

[60] 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, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 6

[61] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic Differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017. 3

[62] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery. 1

[63] Shi Qiu, Saeed Anwar, and Nick Barnes. Dense-resolution network for point cloud classification and segmentation. In *WACV*, pages 3813–3822, 2021. 2

[64] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:1511.06434*, 2015. 1, 5

[65] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, page 9, 2019. 3

[66] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 3

[67] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 5

[68] Franco Ronchetti, Facundo Quiroga, Cesar Estrebou, Laura Lanzarini, and Alejandro Rosete. Lsa64: A dataset of

argentinian sign language. *XX II Congreso Argentino de Ciencias de la Computación (CACIC)*, 2016. 1, 3

[69] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, pages 36479–36494. Curran Associates, Inc., 2022. 4, 1

[70] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Adversarial Training for Multi-Channel Sign Language Production. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020. 3, 6, 7

[71] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 4, 6, 7, 8

[72] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. *International Journal of Computer Vision (IJCV)*, 2021. 4, 6

[73] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Mixed SIGNals: Sign Language Production via a Mixture of Motion Primitives. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 6, 7

[74] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Skeletal Graph Self-Attention: Embedding a Skeleton Inductive Bias into Sign Language Production. *arXiv preprint arXiv:2112.05277*, 2021. 1, 2, 4

[75] 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 (CVPR)*, pages 5141–5151, 2022. 3, 6

[76] Adam Schembri, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. Building the british sign language corpus. *Language Documentation & Conservation*, 7:136–154, 2013. 3

[77] Murray Shanahan. Talking about large language models. *CoRR*, abs/2212.03551, 2022. 2

[78] Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. 1, 4

[79] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable GANs for Pose-Based Human Image Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[80] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020. 1, 3

[81] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 6

[82] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Text2Sign: Towards Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision (IJCV)*, 2020. 2, 3

[83] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 2

[84] Hao Tang, Wei Wang, Dan Xu, Yan Yan, and Nicu Sebe. GestureGAN for Hand Gesture-to-Gesture Translation in the wild. In *Proceedings of the 26th ACM International Conference on Multimedia*, 2018. 1

[85] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. XingGAN for Person Image Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[86] Laia Tarrés, Gerard I. Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró i Nieto. Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2023. 1, 3

[87] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *CoRR*, abs/2211.09085, 2022. 2

[88] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Computer Vision – ECCV 2020*, pages 716–731, Cham, 2020. Springer International Publishing. 1

[89] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, 2023. 2

[90] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing Motion and Content for Video Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 5

[91] Lucas Ventura, Amanda Duarte, and Xavier Giró-i Nieto. Can Everybody Sign Now? Exploring Sign Language Video Generation from 2D Poses. In *ECCV Sign Language Recognition, Translation and Production Workshop*, 2020. 1

[92] U. Von Agris and K.-F. Kraiss. Signum database: Video corpus for signer-independent continuous sign language

recognition. In *Workshop on Representation and Processing of Sign Languages*, pages 243–246, 2010. 3

[93] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating Videos with Scene Dynamics. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 1, 5

[94] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 2, 3, 1, 5

[95] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 1, 5

[96] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1

[97] Dongxu Wei, Xiaowei Xu, Haibin Shen, and Kejie Huang. GAC-GAN: A General Method for Appearance-Controllable Human Video Motion Transfer. *IEEE Transactions on Multimedia*, 2020. 3, 4, 1

[98] Zhenyu Wu, Duc Hoang, Shih-Yao Lin, Yusheng Xie, Liangjian Chen, Yen-Yu Lin, Zhangyang Wang, and Wei Fan. MM-Hand: 3D-Aware Multi-Modal Guided Hand Generative Network for 3D Hand Pose Synthesis. *arXiv preprint arXiv:2010.01158*, 2020. 1

[99] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020. 3

[100] Seunghan Yang, Seungjun Jung, Heekwang Kang, and Changick Kim. The korean sign language dataset for action recognition. In *International conference on multimedia modeling*, pages 532–542. Springer, 2019. 3

[101] Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. Mlslt: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5109–5119, 2022. 3, 6

[102] Lingyun Yu, Jun Yu, Mengyan Li, and Qiang Ling. Multimodal inputs driven talking face generation with spatial–temporal dependency. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1):203–216, 2021. 1

[103] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2019. 1

[104] Jan Zelinka and Jakub Kanis. Neural Sign Language Synthesis: Words Are Our Glosses. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 3

[105] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 4, 5

[106] Sibo Zhang, Jiahong Yuan, Miao Liao, and Liangjun Zhang. Text2video: Text-driven talking-head video synthesis with personalized phoneme - pose dictionary. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2659–2663, 2022. 1

[107] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara Berg. Dance Dance Generation: Motion Transfer for Internet Videos. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019. 3, 4, 1

[108] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 5

[109] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive Pose Attention Transfer for Person Image Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[110] Inge Zwitserlood, Margriet Verlinden, Johan Ros, and Sanny Van Der Schoot. Synthetic Signing for the Deaf: Esign. In *Proceedings of the Conference and Workshop on Assistive Technologies for Vision and Hearing Impairment (CVHI)*, 2004. 1