MS2SL: Multimodal Spoken
Data-Driven Continuous Sign Language Production

Jian Ma[1,3], **Wenguan Wang**[2], **Yi Yang**[2], and **Feng Zheng**[†1]

[1]Southern University of Science and Technology
[2]ReLER, CCAI, Zhejiang University [3]ReLER, University of Technology Sydney
https://hechang25.github.io/MS2SL

Abstract

Sign language understanding has made significant strides; however, there is still no viable solution for generating sign sequences directly from entire spoken content, *e.g.*, text or speech. In this paper, we propose a unified framework for continuous sign language production, easing communication between sign and non-sign language users. In particular, a sequence diffusion model, utilizing embeddings extracted from text or speech, is crafted to generate sign predictions step by step. Moreover, by creating a joint embedding space for text, audio, and sign, we bind these modalities and leverage the semantic consistency among them to provide informative feedback for the model training. This embedding-consistency learning strategy minimizes the reliance on sign triplets and ensures continuous model refinement, even with a missing audio modality. Experiments on How2Sign and PHOENIX14T datasets demonstrate that our model achieves competitive performance in sign language production.

1 Introduction

Sign language, a visual language, combines both manual (hand gestures) and non-manual cues for communication. It is specifically designed for the deaf and hearing-impaired community (Hickok et al., 1996; Armstrong and Wilcox, 2003; Campbell et al., 2008; Zhou et al., 2020). According to the World Federation of the Deaf, there are 70 million deaf people and more than 200 kinds of sign languages in the world (Fenlon and Wilkinson, 2015; Núñez-Marcos et al., 2023). Improvements in sign language production (SLP) can bridge the communication gap between the deaf and hearing (Mehdi and Khan, 2002; Harris et al., 2009; Taskiran et al., 2018; Rastgoo et al., 2021; Kahlon and Singh, 2023; Luo and Yang, 2024).

The challenges primarily arise from **phonological difference** and **data scarcity**. Phonological

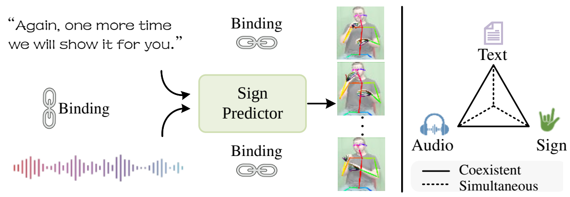

Figure 1: **Illustration of our sign language producer.** 1) We propose a unified, multimodal spoken data-driven framework for SLP that can directly produce sign sequences from spoken text or speech audio. 2) To overcome data scarcity, we train a joint embedding space through the spontaneous alignment of multimodal data. Within this space, we establish a consistency learning strategy to provide feedback signals that boost training.

difference: signs are composed of various manual and non-manual features (Mann et al., 2010), such as hand gestures, facial expressions and limb movements (Liddell and Johnson, 1989; Johnson and Liddell, 2011; Sandler, 2012). The differences in phonological structure and means of expression create challenges in modeling the two languages. Data scarcity: multimodal high-quality sign language datasets are relatively scarce, and some datasets tend to be specific to a particular language or domain, *e.g.*, American sign (Duarte et al., 2021), German weather (Forster et al., 2014; Camgöz et al., 2018). Furthermore, hearing impairments hinder pronunciation (Moeller, 2000; Yoshinaga-Itano, 2003), making it strenuous to collect sign video with aligned audio and usually resulting in the lack of auditory information. Previous researches (Zhang et al.; Camgöz et al., 2017; Hu et al., 2021b,a; Yin et al., 2022) primarily focused on sign language recognition, which identifies sign fragments as the corresponding sign language lexicons (*e.g.*, gloss). Several work (Saunders et al., 2020, 2021a, 2022; Hwang et al., 2021; Walsh et al., 2022) manage the transition from gloss to sign sequences, yet the grammar of gloss can be perplexing for those without sign language training. Saunders et al. (2020, 2021b) can transcribe discrete words or phrases into continuous sign language se-

[†]Corresponding author.

Findings of the Association for Computational Linguistics ACL 2024, pages 7241–7254
August 11-16, 2024

quences. However, directly producing continuous signs from entire spoken sentences still remains more exploration and efforts.

To promote barrier-free communication between signers and speakers, we introduce a Multimodal Spoken Data-Driven Continuous Sign Language Production (MS2SL) framework (Fig. 1). MS2SL can animate sign keypoint sequences from either speech audio or text. In addition, to alleviate data demands, we adopt an embedding-consistency learning (ECL) strategy, which is inherently based on the reciprocity among modalities, to bolster the model training. Specifically, MS2SL initially employs pre-training models like CLIP (text) (Radford et al., 2021) and HuBERT (audio) (Hsu et al., 2021) to extract features from input. Subsequently, we utilize these features, serving as control conditions for the diffusion, to generate sign sequences. The attention mechanism (Vaswani et al., 2017) is employed to model the relationships among conditions, denoising steps, and sign movements. Besides that, ECL does not require the three modalities to coexist in the dataset. By learning a joint embedding space, inspired by ImageBind (Girdhar et al., 2023), ECL tightly binds the properties of different modalities and generates feedback signals to boost the training process. First, we utilize contrastive learning to bind audio and text in the embedding space. Then, we leverage the semantic consistency between co-occurring data to infer and reconstruct the embedding of missing modalitiy. The reconstruction error between the generated signs and groundtruth can be used to iteratively update MS2SL until convergence. ECL can foster cross-learning between different generation streams, allowing training even in the absence of certain modality. Furthermore, the inclusion of audio data not only enriches sample diversity and enhances multimodal comprehension but also assists in accurately capturing the expression and semantic content of sign language. We validate the effectiveness of our method across two prevalent datasets How2Sign (Duarte et al., 2021) and (Camgöz et al., 2018). Experimental results demonstrate that MS2SL achieves SOTA performance, both in terms of semantic consistency and sign accuracy. In conclusion, our primary contributions are outlined as follows:

- We propose MS2SL, a unified diffusion framework for efficient multimodal spoken to sign language production. MS2SL is able to directly convert entire speech or text sentences into corresponding sign keypoints sequences.
- We present an ECL strategy that leverages the intrinsic relations to enhance data utilization.
- We show that joint embedding is suitable for generative tasks that are prone to modality missing.

2 Related Work

Sign Language Understanding. Similar to spoken language, sign language follows specific linguistic rules (Sandler and Lillo-Martin, 2006; Brentari, 2011; Petitto et al., 2016; Sandler, 2017). Existing researches are primarily dedicated to sign language translation (SLT) and recognition. SLT typically involves translating sign language into spoken language (Camgöz et al., 2018; Coster et al., 2022; Camgöz et al., 2020; Lin et al., 2023). Sign language recognition (Adaloglou et al., 2022; Selvaraj et al., 2022) means interpreting and classifying of body movements in videos, covering isolated (Imashev et al., 2020) and continuous signs (Cui et al., 2017; Camgöz et al., 2018, 2020). SLP (Arkushin et al., 2023) is the process of creating sign sequences from spoken text, and can be seen as the reverse process of SLT. These existing studies on SLT and SLP primarily focus on converting between sign videos and discrete glosses, either directly or indirectly. A few of Text2Sign works (Saunders et al., 2020, 2021a,b) are grounded in datasets with relatively homogeneous scenario (Camgöz et al., 2018) and discrete spoken transcriptions.

Diffusion Model. The diffusion model demonstrates exceptional proficiency in various generative tasks (Ho et al., 2020; Choi et al., 2021; Lugmayr et al., 2022; Avrahami et al., 2022). Beyond image generation, diffusion models also perform well in generating sequence data (Yuan et al., 2022; Wu et al., 2023). In recent years, some work has begun to apply diffusion models to SLP. By iteratively updating information, diffusion models can gradually infer the distribution of subsequent data, thereby providing more accurate and coherent results. Ham2Pose (Arkushin et al., 2023) leverages diffusion to animate HamNoSys, a lexicon of sign symbols, into sign keypoint sequences. Though impressive, Ham2Pose can only produce videos with a single sign symbol, falling short in conveying sentences with complete semantics.

Cross-modal Consistency Learning. Deep learning often requires ample labeled data to work properly. However, the cost of collecting sign data is prohibitive and audio data is often lacking. Recent methods enhance model training by applying consistency training to massive unlabeled data (Bachman et al., 2014; Sajjadi et al., 2016; Clark et al., 2018; Miyato et al., 2019). The principle of consistency learning, employing the cyclical duality between different tasks or data as feedback signals to regularize training (He et al., 2016), has its roots in the domain of language translation (Yi et al., 2017; Lu et al., 2017; Zhao et al., 2020). It primarily encompasses inter-task (dual-learning) and intra-task (cycle-consistency learning) varieties. Dual-learning simultaneously trains bidirectional mapping functions between tasks, creating a primal-dual pair where one function's output approximates the input of the inverse function (Yi et al., 2017; Wang et al., 2022; Zhang et al., 2018; Shah et al., 2019; Wang et al., 2019; Zhao et al., 2020; Xie et al., 2020). Cycle-consistency learning is designed to enhance the self-reconstruction capabilities of samples produced intrinsically by the same model (Zhu et al., 2017; Almahairi et al., 2018; Rao et al., 2020; Mathew et al., 2020). However, these methods frequently emphasize the duality between two tasks or modalities, overlooking the interplay and mutual influence among multimodal data within the same task.

Limited studies focus on directly generating sign language sequences from entire spoken sentences. To our best knowledge, we are the pioneers in effecting this conversion. This study harnesses sequential diffusion models to incrementally generate noise predictions, enabling cross-modal sign language generation. With the help of ECL, MS2SL can generate various feedback signals even in the absence of co-occurring ternary data: assessing the reconstruction loss with the signs generated from the reconstructed audio embeddings.

3 Method

Assuming the triplets $(\mathcal{A}, \mathcal{T}, \mathcal{S})$ represent the audio, text, and sign space respectively, our goal is to learn the mapping from text or audio to sign within a unified framework (Fig. 2). Given a training dataset $\mathcal{D} = \{(\boldsymbol{a}, \boldsymbol{t}, \boldsymbol{s}) \in \mathcal{A} \times \mathcal{T} \times \mathcal{S}\}$, MS2SL can realize text-to-sign $\mathcal{T} \mapsto \mathcal{S} : \boldsymbol{s} = G(\boldsymbol{t})$ and audio-to-sign $\mathcal{A} \mapsto \mathcal{S} : \boldsymbol{s} = G(\boldsymbol{a})$, where G is the sign sequence diffusion generator. We initially employ pretrained models CLIP (Radford et al., 2021) and HuBERT (Hsu et al., 2021) to extract features from text $\boldsymbol{t}$ and audio $\boldsymbol{a}$. Next, we employ three encoders E_a, E_t, E_s to encode these features, acquiring their embeddings $\boldsymbol{e}_a$, $\boldsymbol{e}_t$, and $\boldsymbol{e}_s$. Subsequently, drawing on the operating mechanism of diffusion models, we employ a diffusion step encoder E_h and a sign noise encoder E_n to encode step h and noise $\boldsymbol{n}$ to $\boldsymbol{e}_h$ and $\boldsymbol{e}_n$, respectively. Finally, we utilize the generator G to produce the sign sequences: $\hat{\boldsymbol{s}}_t = G(\boldsymbol{e}_t, \boldsymbol{e}_h, \boldsymbol{e}_n)$ and $\hat{\boldsymbol{s}}_a = G(\boldsymbol{e}_a, \boldsymbol{e}_h, \boldsymbol{e}_n)$.

The paucity of co-occurring triplet data renders the direct training of MS2SL a formidable task. To overcome this challenge, we develop a joint embedding space that facilitates the natural alignment of multimodal data. Furthermore, we employ ECL strategy to exploit the reciprocity among modalities within the embedding space, effectively furnishing feedback signals to boost the training.

3.1 Sign Predictor

Cross-linguistic Modeling. MS2SL aims to solve the problem of generating variable-length sequences across modalities. It necessitates phonological modeling between spoken and sign language, associating text and audio to the same target sign sequence. The causal attention mechanism can serve as a potent remedy for this challenging issue. Taking text-to-sign as an example, we first concatenate the embeddings of text $\boldsymbol{e}_t$, denoising step $\boldsymbol{e}_h$ and noise $\boldsymbol{e}_n$. Next, we apply the causal self-attention (Radford et al., 2018) to model the relationship among them. The *mask* in causal attention ensures that the model only processes past and present information, maintaining temporal and logical coherence in the output. As such, the output is computed as: $\text{CausalAtt}[\boldsymbol{e}_t; \boldsymbol{e}_h; \boldsymbol{e}_n]$. During inference, we initiate from the text embedding and produce indices autoregressively, ceasing generation when the model predicts the sequences. Likewise, the concatenated entity of the audio $\boldsymbol{e}_a$, step $\boldsymbol{e}_h$, and noise $\boldsymbol{e}_n$ can also undergo the causal attention to capture the relationship between audio and sign. In causal attention, we adopt the common practice of positional encoding, which can model keypoints and inter-frame context while capturing cross-modal relations. Thus, to simplify the model structure, we does not explicitly design a temporal module. Finally, we employ two fully connected layers to output the sign prediction $\hat{\boldsymbol{s}}_h$ for step h.

Sign Language Production. We apply a diffusion model as the sign generator. Similarly, taking text-to-sign as an example, the diffusion generator G is responsible for the gradually producing a continuous sign sequence $\hat{\boldsymbol{s}}$. Diffusion generator G simulates data distribution through a gradual forward and reversible process (Ho et al., 2020), training by maximizing the evidence lower bound to approxi-

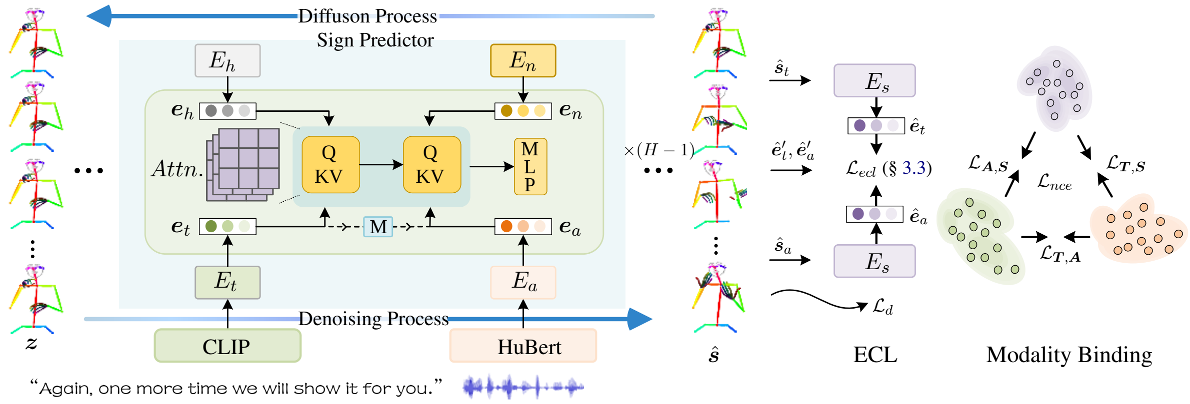

Figure 2: **Overview of our framework for MS2SL.** It includes three key components: sign predictor (§3.1), modality binding (§3.2) and ECL strategy (§3.3). MS2SL directly unifies spoken content from different modalities into a common sign language production framework. The introduction of the joint embedding space and ECL reduces the reliance on co-occurring (text, audio, sign) triplet.

mate target distributions. Diffusion model aims to reconstruct the input from a latent variable. The forward process gradually transforms the input into noise by adding Gaussian noise. The reverse process starts from random noise and progressively removes the noise to recover the original data.

Common training for diffusion models involves independent noise prediction at each forward step h, potentially reducing sequence coherence and consistency. Following (Arkushin et al., 2023), we adopt the holistic training method. We apply a schedule function $\delta_h = 1/\log(h+1)$ ($\delta \in [0,1]$) and a step size $\alpha_h = \delta_h - \delta_{h+1}$. The predicted signs $\hat{\boldsymbol{s}}_h$ at step h, as:

$$\hat{s}_h = \alpha_h p_h + (1-\alpha_h)\hat{s}_{h-1}, \tag{1}$$

where the predicted signs p_h at step h are given as $G(\boldsymbol{t})$. This method utilizes the output from the previous iteration as the input for the subsequent step, gradually reducing the step size as the process continues. Each step combines previous outcomes with current predictions, reducing reliance on the initial noise. We also enhance training robustness by introducing a random noise to $\hat{s}_h$ at each step. Finally, the predicted initial sign $\hat{s}_0$ is outputted. The loss of the diffusion is defined as:

$$\mathcal{L}_d = \alpha_h s_0 + (1-\alpha_h)s_{h+1}. \tag{2}$$

3.2 Modality Binding

MS2SL operates in an aligned embedding space, typically dependent on audio, text, and sign data for tri-modal alignment. However, the difficulty for people with hearing impairments to perceive sound variations poses a challenge in recording these co-occurring triplets. Fortunately, ImageBind (Girdhar et al., 2023) reveals that a model can learn to align modalities in a joint embedding space by employing contrastive learning (Hadsell et al., 2006). Training with (Image, Modality1) and (Image, Modality2) pairs can lead to a spontaneous alignment of Modality1 and Modality2 in embedding space. This alignment allows the model to excel in various tasks without requiring direct training on specific pairs of (Modality1, Modality2).

We extend the findings of ImageBind and construct a joint embedding space for the triplet dataset $(\mathcal{A}, \mathcal{T}, \mathcal{S})$, where MS2SL employs (text, sign) pairs as anchors to establish a cohesive space linking audio, text, and sign. Let's explore a pair of modalities $(\mathcal{T}, \mathcal{S})$ with aligned observations. Given a sign sequence $\boldsymbol{s}$ and its corresponding caption $\boldsymbol{t}$. We first employ pretrained models CLIP (Radford et al., 2021) to extract textual features and encode them into normalized embeddings: $\boldsymbol{e}_t$ and $\boldsymbol{e}_s$. Then, we leverages the paired modalities $(\mathcal{T}, \mathcal{S})$ to align the text with sign. The corresponding encoders are optimized by InfoNCE (Oord et al., 2018) loss $\mathcal{L}_{\mathcal{T},\,\mathcal{S}}$:

$$\mathcal{L}_{\mathcal{T},\,\mathcal{S}} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{e}_t, \boldsymbol{e}_s)/\tau)}{\sum_{m=1}^{M} \exp(\mathrm{sim}(\boldsymbol{e}_t, \boldsymbol{e}_{s_m})/\tau)}. \tag{3}$$

Within the mini-batch, we consider each instance, whose index is not equal to m, as a negative example. This approach aims to draw different embedding pairs closer within their joint embedding space. Similarly, we can also obtain $\mathcal{L}_{\mathcal{A},\,\mathcal{S}}$ and $\mathcal{L}_{\mathcal{T},\,\mathcal{A}}$ for the pairs $(\mathcal{A}, \mathcal{S})$ and $(\mathcal{T}, \mathcal{A})$. Interestingly, we also observe the emergent alignment between modal pairs $(\mathcal{T}, \mathcal{A})$ in our embedding space. This phenomenon can occur when the training is solely based on pairs $(\mathcal{T}, \mathcal{S})$ and $(\mathcal{A}, \mathcal{S})$, a trend that mirrors the findings reported in (Girdhar et al., 2023). Accordingly, MS2SL is designed to mainly leverage modal pairs $(\mathcal{T}, \mathcal{S})$ and $(\mathcal{T}, \mathcal{A})$, circumventing the need for triplet data. In practice, this is achieved by employing a triadic loss:

$$\mathcal{L}_{nce} = \mathcal{L}_{\mathcal{T},\,\mathcal{S}} + \mathcal{L}_{\mathcal{T},\,\mathcal{A}} + \mathcal{L}_{\mathcal{A},\,\mathcal{S}}. \tag{4}$$

As such, the embedding space can not only spontaneously align unseen triples but also be used in reconstructing unobserved modalities in ECL.

3.3 Embedding-consistency Learning

Given a tuple $(\mathcal{A}, \mathcal{T}, \mathcal{S})$, we employ a cyclic approach with the bound joint embedding to generate feedback signals for bidirectional cross-learning, fostering model training. When triplet data is available, the encoders first extract features from their respective modalities. Then, audio and text independently generate predicted sign language sequences $\hat{\boldsymbol{s}}_a$ and $\hat{\boldsymbol{s}}_t$. To fully utilize real data, we calculate ECL loss after 500 epochs of model training. The vanilla model, built on authentic data, guarantees minimal distribution differences between generated pseudo-embeddings and the original dataset. Semantic consistency is calculated using the embeddings $\hat{\boldsymbol{e}}_t$ and $\hat{\boldsymbol{e}}_a$ from encoder E_s, which encodes the two predicted sequences. We can obtain the text-to-sign error $\Delta(\hat{\boldsymbol{e}}_t,\ \boldsymbol{e}_s)$ and the audio-to-sign loss $\Delta(\hat{\boldsymbol{e}}_a,\ \boldsymbol{e}_s)$:

$$\begin{aligned}\Delta(\hat{\boldsymbol{e}}_t,\ \boldsymbol{e}_s) &= \|\hat{\boldsymbol{e}}_t, \boldsymbol{e}_s\|_2,\\ \Delta(\hat{\boldsymbol{e}}_a,\ \boldsymbol{e}_s) &= \|\hat{\boldsymbol{e}}_a, \boldsymbol{e}_s\|_2.\end{aligned} \tag{5}$$

Evaluation scores are derived from comparing the two embeddings $\hat{\boldsymbol{e}}_t$ and $\hat{\boldsymbol{e}}_a$. Both audio and text can receive feedback signals from the generative streams of each other. To compensate for the missing audio modality and ensure smooth processing, we use a mapping network M and text embeddings to generate pseudo audio features. The operation is conducted in the embedding space, thus minimally affecting inference speed. For unpaired natural audios $\mathcal{U}$, we can get the formula:

$$\begin{aligned}\mathcal{L}_{(\mathcal{T},\,\mathcal{A},\,\mathcal{S})} &= \|E_s(G(\boldsymbol{e}_a)) - E_s(G(\boldsymbol{e}_t))\|_2,\\ \mathcal{L}_{(\mathcal{T}',\,\mathcal{S}')} &= \|E_s(M(G(\boldsymbol{e}'_t))) - E_s(G(\boldsymbol{e}'_t))\|_2.\end{aligned} \tag{6}$$

Then our ECL loss is defined as:

$$\mathcal{L}_{ecl} = \mathcal{L}_{(\mathcal{T},\,\mathcal{A},\,\mathcal{S})\in\mathcal{D}} + \mathcal{L}_{(\mathcal{T}',\,\mathcal{S}')\in\mathcal{U}}. \tag{7}$$

MS2SL translates entire spoken sentences into continuous sign language sequences. Overall, our total loss comprises three components, *i.e.*, the diffusion model loss, ECL loss, and joint embedding loss:

$$\mathcal{L} = \lambda_1\mathcal{L}_d + \lambda_2\mathcal{L}_{ecl} + \lambda_3\mathcal{L}_{nce}, \tag{8}$$

where the cofficients are empirically set as $\lambda_1 = \lambda_2 = \lambda_3 = 1$.

3.4 Implementation Details

Training. MS2SL takes speech audio or text as inputs. We utilize pre-trained models for encoding both speech and text, HuBert (Hsu et al., 2021) for speech and CLIP (Radford et al., 2021) for text. We first extract embeddings $\boldsymbol{e}_t, \boldsymbol{e}_a, \boldsymbol{e}_s, \boldsymbol{e}_h, \boldsymbol{e}_n$ through five encoders. We employ keypoints to represent signs, like the 137 human keypoints in How2Sign (Duarte et al., 2021), which are normalized and standardized before being input into the model. $\boldsymbol{e}_t$, $\boldsymbol{e}_a$ and $\boldsymbol{e}_s$ participate in learning the joint embedding space. Concurrently, $\boldsymbol{e}_t, \boldsymbol{e}_a, \boldsymbol{e}_h$ and $\boldsymbol{e}_n$ serve as conditions to control the generation of text-to-sign and audio-to-sign, respectively. Here, we adopt the common practice (Saunders et al., 2021a,b; Arkushin et al., 2023) of using the first sign pose as initial noise. The first 500 epochs skip the audio-to-sign generation flow in the absence of audio. After obtaining a vanilla model, we apply the mapping network M to transform $\boldsymbol{e}_t$ into $\boldsymbol{e}_a$ to continue the training until the model converges. Since PHOENIX (Forster et al., 2014) dataset is in *German* sign language, and our pre-trained model is primarily based on *English*, we utilize the penultimate layer features of CLIP along with MLP to align and transform between German and English. As for ECL, we incorporate cycles among the three modalities, namely audio-to-sign, text-to-sign, and audio-to-text, greatly enhancing the efficiency of data utilization. We adopt the commonly used exponential moving average (Cai et al., 2021) strategy with diffusion parameters (Cai et al., 2021) to ensure smoother, more robust training. For details, please refer to the supplementary.

Inference. The model can perform SLP from audio or text independently. Inference for each modality involves executing the sequence sampling of the diffusion model. Taking text-to-sign as an example, the process starts with CLIP encoding the text into features. These text features are then fed into the sign predictor, which sequentially generates a sequence noise prediction. The completion of this sampling process results in the generation of the desired sign sequence. The process for generating signs from speech is similar. We take the average of twenty generations to mitigate deviation.

Reproducibility. Our method is implemented using PyTorch on 2 RTX 4090 GPUs, with a training time of about 12 hours and an average inference time of 0.3 seconds. Following (Zhang et al., 2023), we remove data with word count exceeding 20.

Methods	How2Sign					PHOENIX14T				
	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
Back-translation	$10.89^{\pm.003}$	$13.32^{\pm.01}$	$16.71^{\pm.06}$	$22.38^{\pm.05}$	$23.23^{\pm.03}$	$20.53^{\pm.01}$	$25.13^{\pm.03}$	$32.81^{\pm.04}$	$44.01^{\pm.02}$	$45.61^{\pm.03}$
PT (Saunders et al., 2020)	$2.01^{\pm.02}$	$3.86^{\pm.04}$	$7.04^{\pm.00}$	$13.69^{\pm.04}$	$13.81^{\pm.03}$	$11.32^{\pm.02}$	$12.91^{\pm.01}$	$19.04^{\pm.05}$	$31.36^{\pm.01}$	$32.46^{\pm.01}$
MOMP (Saunders et al., 2021b)	$2.34^{\pm.04}$	$3.92^{\pm.01}$	$7.63^{\pm.02}$	$13.68^{\pm.06}$	$13.83^{\pm.05}$	$11.19^{\pm.03}$	$13.14^{\pm.02}$	$19.64^{\pm.01}$	$32.22^{\pm.04}$	$32.96^{\pm.02}$
Ham2Pose (Arkushin et al., 2023)	$2.93^{\pm.06}$	$4.07^{\pm.04}$	$7.31^{\pm.02}$	$12.38^{\pm.03}$	$13.29^{\pm.01}$	$11.71^{\pm.03}$	$13.22^{\pm.03}$	$20.16^{\pm.05}$	$33.39^{\pm.00}$	$34.02^{\pm.04}$
T2M-GPT (Zhang et al., 2023)	$3.53^{\pm.03}$	$5.14^{\pm.01}$	$7.92^{\pm.05}$	$12.87^{\pm.05}$	$13.99^{\pm.03}$	$11.66^{\pm.02}$	$13.35^{\pm.07}$	$21.19^{\pm.00}$	$35.24^{\pm.02}$	$35.44^{\pm.03}$
MS2SL *w/o* ECL	$3.76^{\pm.02}$	$6.03^{\pm.02}$	$8.05^{\pm.04}$	$14.51^{\pm.05}$	$15.10^{\pm.06}$	$12.03^{\pm.02}$	$14.32^{\pm.04}$	$21.72^{\pm.03}$	$35.36^{\pm.06}$	$35.68^{\pm.08}$
MS2SL-T2S	**4.26**$^{\pm.04}$	**6.84**$^{\pm.02}$	**9.17**$^{\pm.05}$	**14.67**$^{\pm.03}$	**16.38**$^{\pm.03}$	**12.77**$^{\pm.06}$	**15.81**$^{\pm.07}$	**22.04**$^{\pm.03}$	**36.41**$^{\pm.01}$	**36.63**$^{\pm.03}$

Table 1: **Comparisons of text-to-sign with the state-of-the-art methods (§4.2) on How2Sign (Duarte et al., 2021) and PHOENIX14T (Camgöz et al., 2018).** For each metric, we repeat the evaluation 20 times and report the average. Red and Blue indicate the best and the second best result, respectively.

4 Experiments

We evaluate the effectiveness of MS2SL under text-to-sign and audio-to-sign settings.

4.1 Experimental Setup

Datasets. We conduct experiments on two continuous sign language datasets:

- How2Sign (Duarte et al., 2021) is a challenging multimodal American sign language dataset with a 16k-word vocabulary and comprehensive annotations. It includes $1,176$ entries with audio and has train/dev/test splits of $31165/1741/2357$.
- PHOENIX14T (Camgöz et al., 2018), a widely applied German weather sign language dataset, contains $2,887$ words, $1,066$ sign annotations, with train/dev/test splits of $7096/519/642$.

Evaluation Metrics. Following (Saunders et al., 2020), we adopt back-translation approach for evaluating, *i.e.*, we leverage the cutting-edge SLT model (Camgöz et al., 2020) to ingeniously translate back from generated signs to text. Subsequently, we calculate BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) scores, which are commonly used metrics for SLP and machine translation. We apply ROUGE-L F1-Score and report BLEU-1 to BLEU-4 for translation performance at different phrase lengths.

Competitors. For text-to-sign generation stream, we consider four SOTA competitors:

- Ham2Pose (Arkushin et al., 2023), which employs transformer and diffusion model, animates HamNoSys (a sign notation) into sign poses.
- T2M-GPT (Zhang et al., 2023) combines VQ-VAE (van den Oord et al., 2017) and CLIP (Radford et al., 2021) for motion generation.
- PT (Saunders et al., 2020) translates discrete spoken sentences into sign sequences.
- MOMP (Saunders et al., 2021b) divides SLP into two sub-tasks: latent sign representation and animation imitation.

As for the audio-to-sign stream, since there are not specific methods, we extend MS2SL to multiple implementations for a thorough evaluation, including audio-to-sign, audio-to-text-to-sign, and text-to-audio-to-sign. For audio-to-text-sign, we apply WeNet (Yao et al., 2021) to translate audio into text, followed by the generation of signs. Conversely, for text-to-audio-to-sign, we employ DeepVoice (Gibiansky et al., 2017) to convert text into audio for subsequent sign generation.

4.2 Comparison to State-of-the-art

Quantitative Results. We present the comparative analysis results in Table 1 on How2Sign and PHOENIX14T test set. MS2SL demonstrates impressive gains against the four robust methods, establishing a new benchmark for SOTA performance. In the generation of text-to-sign, our approach yields a ROUGE of 14.67, marking a notable increase of 2.39 over its counterpart (T2M-GPT, which has a 13.99 ROUGE). Furthermore, MS2SL combined with ECL surpasses the standalone by 1.28. How2Sign (Duarte et al., 2021) and PHOENIX14T (Camgöz et al., 2018) are datasets of different scales, demonstrating the robustness of our method and the burgeoning potential of diffusion models in generating long sign sequences.

Table 2 reports the audio-to-sign results on How2Sign, noting that PHOENIX14T is not included here due to the absence of audio data. Our method significantly enhance performance, achieving notable improvements (*i.e.*, BLEU-1 increase from 9.49 to 11.77, ROUGE from 9.60 to 12.16). The ECL strategy also enhances ROUGE by 1.12. Considering the scarcity of audio modality data, this achievement is particularly noteworthy and shows its real-world applicability. We can also conclude that it is difficult to obtain a well-performing model by training solely with the limited audio data in How2Sign. This also highlights the urgency of utilizing non-co-occurring triplets.

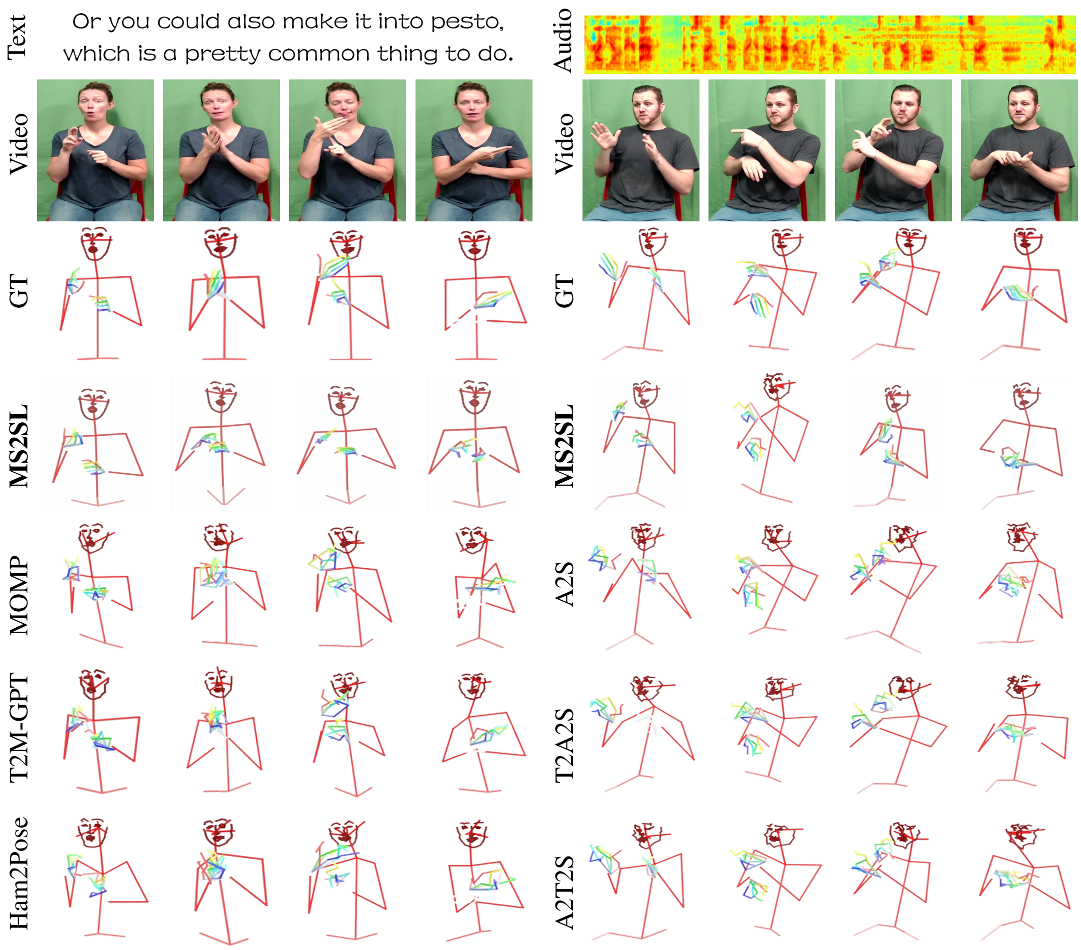

Figure 3: **Results examples (§4.2): Left column:** text-to-sign generation stream, **right column:** audio-to-sign generation stream. Under given conditions, our MS2SL can generate signs that are more semantically consistent with the spoken description and have more precise keypoints.

Methods	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
T2A2S (§4.1)	$0.98^{\pm.04}$	$1.32^{\pm.02}$	$3.71^{\pm.02}$	$8.38^{\pm.01}$	$8.52^{\pm.00}$
A2T2S (§4.1)	$1.02^{\pm.02}$	$1.47^{\pm.01}$	$4.66^{\pm.05}$	$9.49^{\pm.08}$	$9.60^{\pm.04}$
MS2SL *w/o* ECL	$1.24^{\pm.07}$	$1.63^{\pm.03}$	$4.71^{\pm.01}$	$10.59^{\pm.01}$	$11.04^{\pm.03}$
MS2SL-A2S	**$1.67^{\pm.01}$**	**$1.94^{\pm.03}$**	**$5.90^{\pm.02}$**	**$11.77^{\pm.05}$**	**$12.16^{\pm.01}$**

Table 2: **Audio-to-Sign results on How2Sign (§4.2).**

Methods	How2Sign	PHOENIX14T
PT (Saunders et al., 2020)	1.29	1.54
Ham2Pose (Arkushin et al., 2023)	1.97	1.73
A2T2S (§4.1)	1.87	2.09
T2M-GPT (Zhang et al., 2023)	2.19	2.20
MS2SL	**2.65**	**3.21**

Table 3: **User study (§4.2).**

Qualitative Comparison. Fig. 3 presents visual results on How2Sign (Duarte et al., 2021). It demonstrates that our method can produce signs that are more closely aligned with their semantic meaning. After meticulous examination, it is evident that MS2SL surpasses other models in generating actions with smoother transitions, heightens expressiveness, greater diversity, and superior adherence to physical constraints. Some noise and jitter are noted in the audio-to-sign generation stream. The main reason is that our method focuses on translating complete spoken content into sign sequences, whereas previous studies (Saunders et al., 2020, 2021a; Arkushin et al., 2023) target the creation of discrete lexical symbol or phrase. The challenge of training models to convey extended semantic content and long sequences often leads to incoherent movements during sign generation.

User Study. Given the challenge of finding sign language experts, who require extensive training, we conduct a user study with 10 hearing volunteers. We ask the volunteers to compare sign sequences generated by different methods. We slow down sign sequence playback for easier comparison by volunteers. Volunteers select the sequence closer to the ground truth and assign a score. Our scoring range is from 1 to 5, with higher scores indicating closer proximity to the ground truth. Most participants report that the sign sequences generated by MS2SL are smoother and more accurate (Table 3). User feedback highlight the advantages of MS2SL in terms of expression clarity and pose accuracy.

Methods	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
Audio	$0.98^{\pm.04}$	$1.32^{\pm.02}$	$3.71^{\pm.02}$	$8.38^{\pm.01}$	$8.52^{\pm.00}$
Text	$1.74^{\pm.00}$	$2.41^{\pm.02}$	$3.43^{\pm.07}$	$8.62^{\pm.03}$	$9.57^{\pm.01}$
T2A2S	$1.85^{\pm.0}$	$2.35^{\pm.03}$	$4.26^{\pm.02}$	$8.52^{\pm.08}$	$9.28^{\pm.03}$
MS2SL	$4.26^{\pm.04}$	$6.84^{\pm.02}$	$9.17^{\pm.05}$	$14.67^{\pm.03}$	$16.38^{\pm.03}$

(a) data from different modalities

Methods	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
$0k$	$3.76^{\pm.06}$	$6.03^{\pm.02}$	$8.05^{\pm.05}$	$14.51^{\pm.04}$	$15.10^{\pm.02}$
$5k$	$3.79^{\pm.06}$	$6.23^{\pm.02}$	$8.17^{\pm.05}$	$14.62^{\pm.04}$	$15.56^{\pm.02}$
$10k$	$3.82^{\pm.03}$	$6.37^{\pm.03}$	$8.31^{\pm.02}$	$14.57^{\pm.06}$	$15.87^{\pm.00}$
$15k$	$4.26^{\pm.04}$	$6.84^{\pm.02}$	$9.17^{\pm.05}$	$14.67^{\pm.03}$	$16.38^{\pm.03}$

(b) embedding consistency learning

Steps	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
0	$0.62^{\pm.02}$	$2.08^{\pm.03}$	$4.16^{\pm.07}$	$9.57^{\pm.03}$	$9.72^{\pm.05}$
5	$1.09^{\pm.04}$	$2.42^{\pm.06}$	$5.24^{\pm.04}$	$10.44^{\pm.00}$	$10.90^{\pm.01}$
10	$4.26^{\pm.04}$	$6.84^{\pm.02}$	$9.17^{\pm.05}$	$14.67^{\pm.03}$	$16.38^{\pm.03}$
15	$4.04^{\pm.01}$	$6.23^{\pm.04}$	$9.58^{\pm.0}$	$15.26^{\pm.02}$	$17.33^{\pm.01}$
20	$4.87^{\pm.01}$	$6.66^{\pm.04}$	$9.67^{\pm.0}$	$15.45^{\pm.02}$	$17.24^{\pm.01}$

(c) diffusion model

Pre-trained	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
WavLM (Chen et al., 2022)	$1.63^{\pm.06}$	$1.79^{\pm.02}$	$6.12^{\pm.01}$	$10.94^{\pm.00}$	$11.43^{\pm.02}$
HuBert (Hsu et al., 2021)	$1.67^{\pm.07}$	$1.94^{\pm.04}$	$5.90^{\pm.02}$	$11.77^{\pm.06}$	$12.16^{\pm.01}$
CLIP (Radford et al., 2021)	$4.26^{\pm.04}$	$6.84^{\pm.02}$	$9.17^{\pm.05}$	$14.67^{\pm.03}$	$16.38^{\pm.03}$
BERT (Devlin et al., 2019)	$4.11^{\pm.04}$	$6.91^{\pm.02}$	$10.27^{\pm.01}$	$13.37^{\pm.05}$	$16.52^{\pm.06}$

(d) different pretrained models

Table 4: **A set of ablation studies (§4.3).** All experiments employ the same network and structure, with slight variations arising due to different inputs. We report the results of text-to-sign generation by default.

4.3 Ablation Study

We conduct careful profiling of the impact of each module within MS2SL on How2Sign.

Data in Different Modalities. We primarily conduct four experiments: audio-to-sign, text-to-sign, text-to-audio-to-sign, and MS2SL, to compare and analyze the role of different modalities. As shown in Table 4a, although direct generation from audio-to-sign and text-to-sign can yield appropriate results, MS2SL significantly outperforms them. Removal of text data leads to a 6.29 decrease in BLEU-1, highlighting its crucial role. The mediating role of text leads to an increase 0.76 in ROUGE. Multimodal data yields superior results compared to its unimodal counterpart, enriching the learning process with more diverse information.

Embedding Consistency Learning. We investigate the impact of the cyclical consistency training presented in § 3.3, and the results are illustrated in Table 4b. We note that common training method performs comparably to baseline models, while cyclical consistency boosts model performance akin to adding substantial training data. Compared to the alternative only with single modality, MS2SL approach shows a 1.12 increase in BLEU-2 and a 1.28 increase in ROUGE, demonstrating the synergistic effect of integrating data from multiple modalities. We further pay particular attention to the impact of dataset size. We also observe a direct correlation between dataset size and model accuracy. For smaller datasets (under $10k$ samples), the accuracy plateau around 15.5. Several insights can be drawn: i) Performances improve as more training data is used. ii) Over $10k$ unpaired data entries, the signs might be of good quality, but the model cannot further improve on a large scale, possibly due to the scarcity of audio. This trend shows that more data notably improves sequence generation, even without clear semantic boundaries.

Diffusion Model. As shown in Table 4c, implementing the diffusion model lead to a significant enhancement. The quality metrics, such as BLEU-1 and ROUGE, improved by 5.1 and 6.66, respectively, compared to non-diffusion model approach. Our study explores denoising steps ranging from 5 to 20, revealing a discernible trade-off between generation quality and computational efficiency. Compared to a fixed 10-step denoising process, the 20-step process unsteadily improve 0.78 in BLEU-1 by approximately 5.3% with a disproportionate increase in computational load. Thus, in this paper, 10 is set as the default number of denoising steps.

Pre-trained Models. We select four widely used models, including, CLIP (text), BERT (text) (Devlin et al., 2019), HuBert (audio) and WavLM (audio) (Chen et al., 2022), to assess their impact on performance. As shown in Table 4d, for audio-to-sign generation, the impact of HuBert and WavLM on performance is minor, with negligible differences observed between the two pre-trained models. GPT outperforms CLIP models in text-related tasks, with a slight improvement of up to 0.14 in ROUGE. This may be because BERT focuses on natural language processing, leading to enhanced text understanding capabilities.

5 Conclusion

We explore a unified framework that combines diffusion and pretrained models to generate sign language from spoken depictions. We surpass other competitors and solidify this classic framework as a highly competitive method for SLP. MS2SL effectively handles diverse modalities of data for analysis and decoupling. Despite its advancements, our model struggles with maintaining contextual flow in generation, and MS2SL cannot handle lengthy data, which is a future focus. Our research pioneers direct sign language generation from speech, offering some insights to advance the community.

Limitations

Despite significant advancements, our method still faces key technical limitations. First, the complexity and fluidity of authentic sign language are challenging to fully capture and reproduce, as it involves not just hand movements but also facial expressions, body language, and the speed of gestures. Moreover, converting text or speech into sign language involves complex natural language processing challenges, especially in handling grammar and semantics. Lastly, MS2SL struggles to effectively generate long sequences of key movements, limiting the coherence and completeness of sign language expression. These limitations indicate that, while the potential of sign language generation technology is immense, significant technical barriers still need to be overcome to achieve comprehensive and precise sign language communication. These are also the directions we are committed to addressing in the future.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 62122035) and (No. 62372405), as well as by the CCF-Tencent Open Fund.

References

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th. Papadopoulos, Vassia Zacharopoulou, George J. Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. 2022. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE Trans. Multim.*, 24:1750–1762.

Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron C. Courville. 2018. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *ICML*.

Rotem Shalev Arkushin, Amit Moryossef, and Ohad Fried. 2023. Ham2pose: Animating sign language notation into pose sequences. In *CVPR*.

David F Armstrong and Sherman Wilcox. 2003. Origins of sign languages. *Deaf studies language and education*, pages 305–318.

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *CVPR*.

Philip Bachman, Ouais Alsharif, and Doina Precup. 2014. Learning with pseudo-ensembles. In *NeurIPS*.

Diane Brentari. 2011. Sign language phonology. *The handbook of phonological theory*, pages 691–721.

Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Zhuowen Tu, and Stefano Soatto. 2021. Exponential moving average normalization for self-supervised and semi-supervised learning. In *CVPR*.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. Subunets: End-to-end hand shape and continuous sign language recognition. In *ICCV*.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *CVPR*.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*.

Ruth Campbell, Mairéad MacSweeney, and Dafydd Waters. 2008. Sign language and the brain: a review. *Journal of deaf studies and deaf education*, 13(1):3–20.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. 2021. ILVR: conditioning method for denoising diffusion probabilistic models. In *ICCV*.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc V. Le. 2018. Semi-supervised sequence modeling with cross-view training. In *EMNLP*.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2022. Machine translation from signed to spoken languages: State of the art and challenges. *CoRR*.

Runpeng Cui, Hu Liu, and Changshui Zhang. 2017. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Amanda Cardoso Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giró-i-Nieto. 2021. How2sign: A large-scale multimodal dataset for continuous american sign language. In *CVPR*.

Jordan Fenlon and Erin Wilkinson. 2015. Sign languages in the world. *Sociolinguistics and Deaf communities*, 1:5.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *LREC*.

Andrew Gibiansky, Sercan Ömer Arik, Gregory Frederick Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep voice 2: Multi-speaker neural text-to-speech. In *NeurIPS*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. In *CVPR*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*.

Raychelle Harris, Heidi M Holmes, and Donna M Mertens. 2009. Research ethics in sign language communities. *Sign Language Studies*, 9(2):104–131.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *NeurIPS*.

Gregory Hickok, Ursula Bellugi, and Edward S Klima. 1996. The neurobiology of sign language and its implications for the neural basis of language. *Nature*, 381(6584):699–702.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *TASLP*, 29:3451–3460.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021a. Signbert: Pretraining of hand-model-aware representation for sign language recognition. In *ICCV*.

Hezhen Hu, Wengang Zhou, and Houqiang Li. 2021b. Hand-model-aware sign language recognition. In *AAAI*.

Eui Jun Hwang, Jung-Ho Kim, and Jong C. Park. 2021. Non-autoregressive sign language production with gaussian space. In *BMVC*.

Alfarabi Imashev, Medet Mukushev, Vadim Kimmelman, and Anara Sandygulova. 2020. A dataset for linguistic understanding, visual evaluation, and recognition of sign languages: The K-RSL. In *CoNLL*.

Robert E Johnson and Scott K Liddell. 2011. Toward a phonetic representation of signs: Sequentiality and contrast. *Sign Language Studies*, 11(2):241–274.

Navroz Kaur Kahlon and Williamjeet Singh. 2023. Machine translation from text to sign language: a systematic review. *Univers. Access Inf. Soc.*, 22(1):1–35.

Scott K Liddell and Robert E Johnson. 1989. American sign language: The phonological base. *Sign language studies*, 64(1):195–277.

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

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. In *ACL*.

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *NeurIPS*.

Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*.

Yawei Luo and Yi Yang. 2024. Large language model and domain-specific model collaboration for smart education. *Frontiers Inf. Technol. Electron. Eng.*, 25(3):333–341.

Wolfgang Mann, Chloe R Marshall, Kathryn Mason, and Gary Morgan. 2010. The acquisition of sign language: The impact of phonetic complexity on phonology. *Language Learning and Development*, 6(1):60–86.

Shawn Mathew, Saad Nadeem, Sruti Kumari, and Arie E. Kaufman. 2020. Augmenting colonoscopy using extended and directional cyclegan for lossy image translation. In *CVPR*.

Syed Atif Mehdi and Yasir Niaz Khan. 2002. Sign language recognition using sensor gloves. In *ICONIP*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2019. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1979–1993.

Mary Pat Moeller. 2000. Early intervention and language development in children who are deaf and hard of hearing. *Pediatrics*, 106(3):e43–e43.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *ICML*.

Adrián Núñez-Marcos, Olatz Perez-de-Viñaspre, and Gorka Labaka. 2023. A survey on sign language machine translation. *Expert Syst. Appl.*, 213(Part):118993.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Laura-Ann Petitto, Clifton Langdon, Adam Stone, Diana Andriola, Geo Kartheiser, and Casey Cochran. 2016. Visual sign phonology: Insights into human reading and language from a natural soundless phonology. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(6):366–381.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Kanishka Rao, Chris Harris, Alex Irpan, Sergey Levine, Julian Ibarz, and Mohi Khansari. 2020. Rl-cyclegan: Reinforcement learning aware simulation-to-real. In *CVPR*.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Syst. Appl.*, 164:113794.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NeurIPS*.

Wendy Sandler. 2012. The phonological organization of sign languages. *Language and linguistics compass*, 6(3):162–182.

Wendy Sandler. 2017. The challenge of sign language phonology. *Annual review of Linguistics*, 3:43–63.

Wendy Sandler and Diane Carolyn Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge University Press.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. In *ECCV*.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2021a. Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. *IJCV*, 129(7):2113–2135.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2021b. Mixed signals: Sign language production via a mixture of motion primitives. In *ICCV*.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *CVPR*.

Prem Selvaraj, Gokul N. C., Pratyush Kumar, and Mitesh M. Khapra. 2022. Openhands: Making sign language recognition accessible with pose-based pretrained models across languages. In *ACL*.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *CVPR*.

Murat Taskiran, Mehmet Killioglu, and Nihan Kahraman. 2018. A real-time system for recognition of american sign language by using deep learning. In *TSP*.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *NeurIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Harry Walsh, Ben Saunders, and Richard Bowden. 2022. Changing the representation: Examining language representation for neural sign language production. *CoRR*.

Hanqing Wang, Wei Liang, Jianbing Shen, Luc Van Gool, and Wenguan Wang. 2022. Counterfactual cycle-consistent learning for instruction following and generation in vision-language navigation. In *CVPR*.

Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2019. Multi-agent dual learning. In *ICLR*.

Tong Wu, Zhihao Fan, Xiao Liu, Yeyun Gong, Yelong Shen, Jian Jiao, Hai-Tao Zheng, Juntao Li, Zhongyu Wei, Jian Guo, Nan Duan, and Weizhu Chen. 2023. Ar-diffusion: Auto-regressive diffusion model for text generation. In *NeurIPS*.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *NeurIPS*.

Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. 2021. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In *Interspeech*.

Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*.

Qifang Yin, Wenqi Tao, Xiaolong Liu, and Yusheng Hong. 2022. Neural sign language translation with sf-transformer. In *ICIAI*.

Christine Yoshinaga-Itano. 2003. From screening to early identification and intervention: Discovering predictors to successful outcomes for children with significant hearing loss. *Journal of deaf studies and deaf education*, 8(1):11–30.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. 2022. Seqdiffuseq: Text diffusion with encoder-decoder transformers. *CoRR*.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Shan Ying. 2023. Generating human motion from textual descriptions with discrete representations. In *CVPR*.

Jihai Zhang, Wengang Zhou, Chao Xie, Junfu Pu, and Houqiang Li. Chinese sign language recognition with adaptive HMM. In *ICME*.

Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *CVPR*.

Zhibing Zhao, Yingce Xia, Tao Qin, Lirong Xia, and Tie-Yan Liu. 2020. Dual learning: Theoretical study and an algorithmic extension. In *ACML*.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2020. Spatial-temporal multi-cue network for continuous sign language recognition. In *AAAI*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.

Supplementary Material

In the appendix, we provide the following components that offer a more comprehensive understanding of our method:

- §B: More Experimental Results.
- §A: Architecture Details.
- §C: Impacts.

We employ GPT-3.5 to refine and enhance our writing. We are immensely grateful for the substantial assistance provided by GPT.

A More Experimental Results

We conducte multiple experiments and report the results on PhoenixT (Camgöz et al., 2018), and we does not conduct comparative experiments for the audio-to-sign generation stream due to the absence of any audio data in PhoenixT. As shown in Table 7 and 8, we report the diagnostic results on PhoenixT. After integrating the diffusion process into MS2SL, we note modest enhancements. In particular, BLEU-1 score shows a notable improvement, rising by 1.88, and ROUGE score experience a increment of 2.11 (Table 7).

Methods	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
0	$10.19.^{\pm.01}$	$12.58^{\pm.04}$	$18.48^{\pm.05}$	$31.92^{\pm.04}$	$33.80^{\pm.01}$
5	$10.91^{\pm.03}$	$13.11^{\pm.02}$	$21.47^{\pm.05}$	$33.8^{\pm.06}$	$35.91^{\pm.00}$
10	$12.77^{\pm.06}$	$15.81^{\pm.07}$	$22.04^{\pm.03}$	$36.41^{\pm.01}$	$36.63^{\pm.03}$
15	$12.97^{\pm.02}$	$15.06^{\pm.01}$	$22.25^{\pm.05}$	$36.73^{\pm.02}$	$37.10^{\pm.02}$

Table 7: **Denoising steps.**

In the comparison of pre-trained models (Table 8), the conclusion is similar to that with How2Sign, indicating no significant differences among various text pre-training models. This is due to the relatively small dataset and vocabulary size of PhoenixT, for which the current models are sufficiently.

Methods	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
GT	$20.53^{\pm.01}$	$25.13^{\pm.03}$	$32.81^{\pm.04}$	$44.01^{\pm.02}$	$45.61^{\pm.03}$
CLIP (Radford et al., 2021)	$12.77^{\pm.06}$	$15.81^{\pm.07}$	$22.04^{\pm.03}$	$36.41^{\pm.01}$	$36.63^{\pm.03}$
Bert (Devlin et al., 2019)	$12.52^{\pm.02}$	$15.76^{\pm.04}$	$22.39^{\pm.03}$	$37.13^{\pm.02}$	$36.45^{\pm.05}$

Table 8: **Different text pre-trained models.**

B Architecture Details

The sign predictor, designed for predicting noise at each step h in the diffusion process (Nichol and Dhariwal, 2021; Arkushin et al., 2023), boasts a streamlined network architecture with several specialized modules. Table 9 details the parameter configurations of each module in MS2SL.

Encoders. We employ a total of five encoders to process different types of input content. Each encoder consists of two attention layers and a Multi-Layer Perceptron (MLP). Attention mechanisms (Radford et al., 2018) in each encoder enable the model to focus on the most relevant features of input data, enhancing its ability to extract and learn complex patterns. MLP further processes those focused information to generate embeddings e_t, e_a, e_s, e_h and e_n, introducing nonlinear transformations to add depth to the analysis and enabling the extraction of higher-level features.

Producer. The producer is a central component of the model, responsible for synthesizing and outputting the final sign predictions. MS2SL utilizes the attention mechanism to learn the relationships between different input content, gathered and processed by the encoders. We utilize six multi-head attention blocks. Finally, we also use an MLP to transform the predicted features into coordinates for 137 sign keypoints in How2Sign (Duarte et al., 2021) and PHOENIX14T (Camgöz et al., 2018).

We also designed a length predictor to forecast the length of the generated sign language sequences. By accurately predicting the sequence length, the length predictor helps maintain the coherence and consistency of the model's outputs, ensuring they are accurate not only in content but also in their temporal unfolding. To reduce the overall parameters of the model, we employed separate predictors for estimating the length of the input text and audio, respectively.

C Impacts

Sign language production technology has significant impacts in both social and technological areas. Socially, it greatly enhances accessible communication, improving information access and interaction for deaf and hard of hearing individuals, especially in daily life, education, and work environments. It can foster social inclusiveness, aiding in the dismantling of communication barriers and facilitating the integration of the deaf community into broader society. SLP also serves as an educational tool, aiding deaf students in better understanding and absorbing information and facilitating the learning of sign language for hearing individuals. Technologically, the advancement of SLP drives progress in image recognition, natural language processing, and machine learning. This involves tackling challenges such as multimodal learning, text and audio comprehension, content generation, and data scarcity simultaneously. We conduct cyclic consistency learning on a joint embedding space, pro-

Module	Text Encoder	Audio Encoder	Sign Encoder	Step Encoder	Noise Encoder
Input	text	audio	sign keypoints	step number	sign noise
Feature extraction	CLIP (Radford et al., 2021)	Hubert (Hsu et al., 2021)	Positional Encoding	nn.Embedding	Positional Encoding
Embedding Generation	Attention Blk $\times 2$ MLP	Attention Blk $\times 2$ MLP	Attention Blk $\times 2$ MLP	Attention Blk $\times 2$ MLP	Attention Blk $\times 2$ MLP
Embedding Fusion	Length prediction (MLP)		-		
	Concatenation				
Sign Prediction	Attention Blk $\times 6$ MLP				

Table 9: **Network architecture of the sign predictor (§B).**

viding effective insights for niche domains. It aslo poses some potential risks, including insufficient accuracy, cultural nuances, and misinterpretations.