Word-Conditioned 3D American Sign Language Motion Generation

Lu Dong, Xiao Wang, Ifeoma Nwogu

University at Buffalo
The State University of New York

Motivation

- Sign words are the building blocks of any sign language.
- Most continuous sign language generation/production are limited to closed-set and struggle with unseen words or phrases. However, we observe that new American Sign Language (ASL) signers can construct a wide range of signs using a fixed set of sign words. This highlights the importance of sign-word synthesis.
- We present **wSignGen**, a word-conditioned 3D American Sign Language generation model, dedicated to synthesizing realistic and grammatically accurate motion sequence for sign words.

Sign Word: **Finish**

Project Page

wSignGen Overview

- **Problem Formulation:**

Our goal is to generate 3D SMPLX-based motion sequences that match the meanings of sign language words, based on either input words or images.

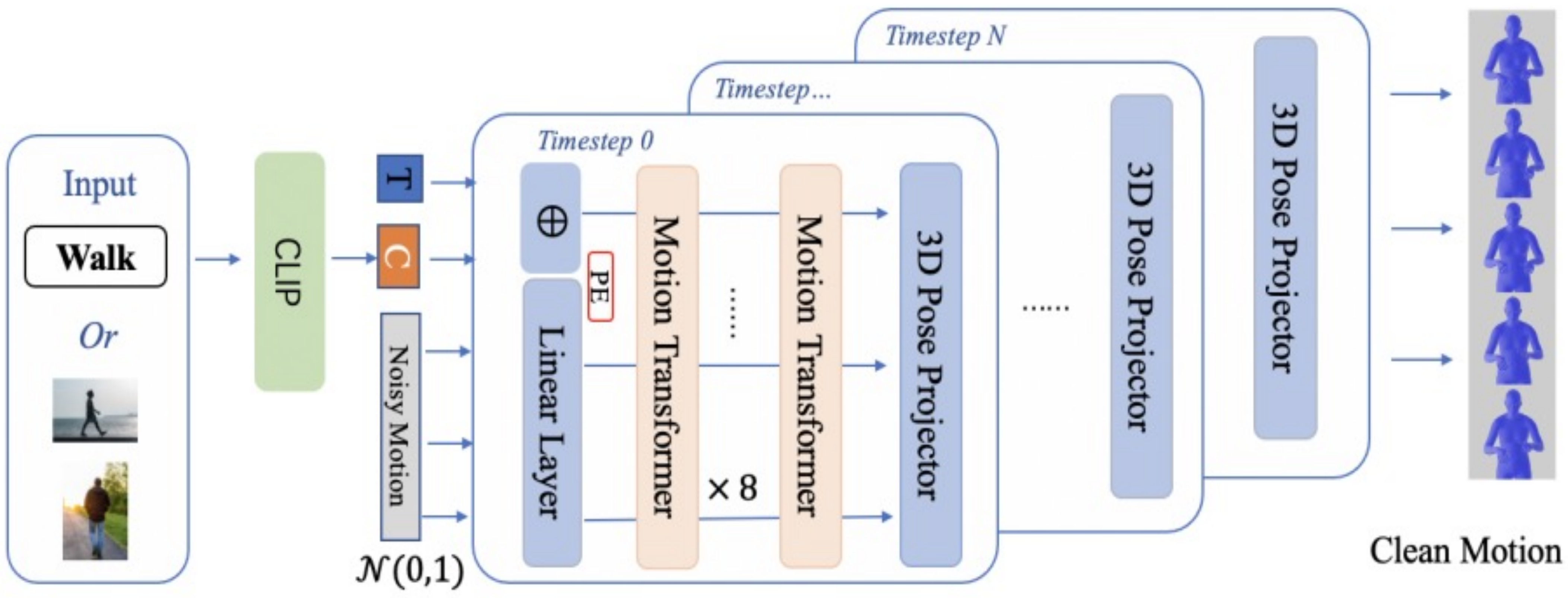
- **Conditioned Word or Image**
- **Diffusion Process**
  - **Training losses**

$$\mathcal{L}_{base} = \mathrm{E}_{X_0 \sim q(X_0|c), t \in [1,t]} \left[\|X_0 - G_c(X_t, t)\|_2^2 \right]$$

$$\mathcal{L}_{vel} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left\| (x_0^{i+1} - x_0^i) - (\hat{x}_0^{i+1} - \hat{x}_0^i) \right\|_2^2$$

$$\mathcal{L} = \mathcal{L}_{base} + \mathcal{L}_{vel}$$

  - **Sampling**



- **wSignGen** not only performs word-conditioned generation but also offers two key advantages:
  - **Image-based generation**, which is especially useful for children learning sign language who may not yet be able to read.
  - The ability to **generalize to unseen synonyms**, allowing for more flexible and comprehensive sign language synthesis.

Experiments

- **Dataset**:
  - We curated a 3D SMPLX-based dataset from the sign recognition video dataset WLASL.
  - Top 30 words denoted as $\mathrm{ASL3D_S}$ for our scalability evaluation.
  - Next, we used the Hand4whole model to extract SMPLX features, creating our SMPLX-based ASL3D Dataset.

- **Quantitative Results**

Table 1: **Comparison of CVAE Baseline and our Diffusion Model** We compare a motion generation baseline algorithm with our proposed method using the curated datasets. Notation Keys: →: implies that motions are better when the metric is closer to those computed for GT^{train} and GT^{test}; "Acc.": accuracy; "Div.": diversity; "Mul.": multimodality; "Gen": Generation. Gen^{train} and Gen^{test} are generated from the same model, and we report them separately to compare with the original training and testing data distribution on FID, Div., and Mul. metrics.

$\mathrm{ASL3D_S}$	Acc. ↑	FID ↓	Div.→	Mul.→	ASL3D	Acc.↑	FID↓	Div.→	Mul.→
Original Data (no Generative Process)									
GT^{train}	1.0	-	30.001	9.921	GT^{train}	1.0	-	34.565	13.256
GT^{test}	0.897	-	26.252	11.180	GT^{test}	0.765	-	30.599	12.289
CVAE Baseline (ACTOR^{+})									
Gen^{train}	0.884	75.243	24.566	8.250	Gen^{train}	0.515	126.830	25.732	16.500
Gen^{test}	-	65.285	24.187	6.600	Gen^{test}	-	100.147	25.393	12.289
wSignGen (Our Model)									
Gen^{train}	1.0	5.348	29.592	8.855	Gen^{train}	1.0	7.339	33.927	11.417
Gen^{test}	-	40.834	29.278	6.494	Gen^{test}	-	37.873	33.608	8.538

- **Human Evaluation Results**

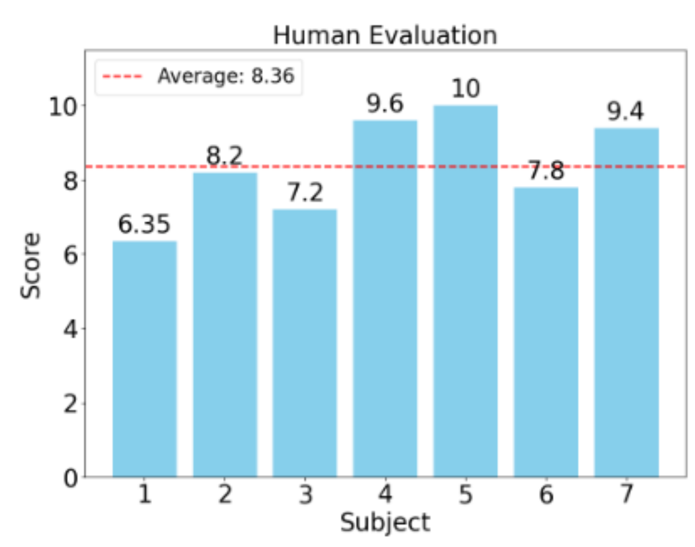
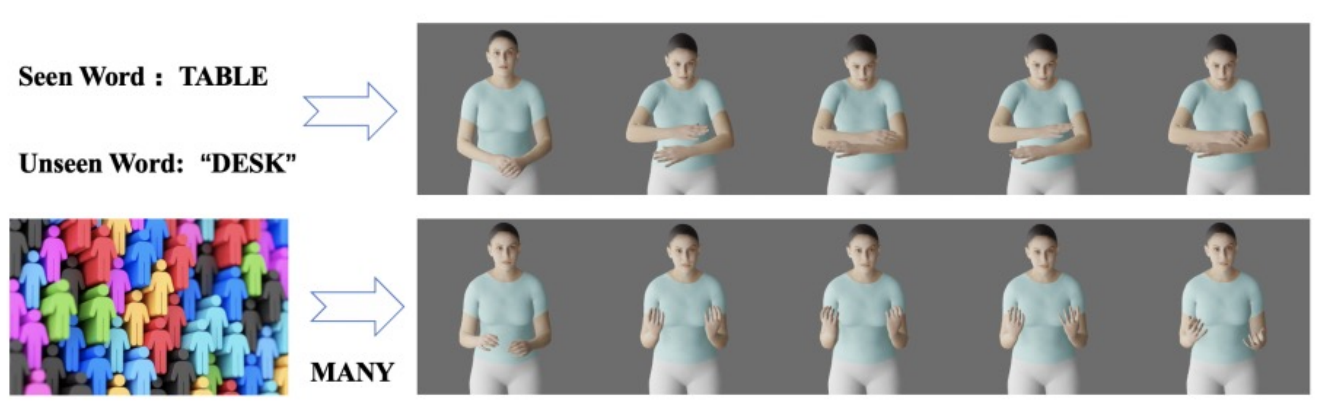


Figure 3: Human Evaluation Results

- **Evaluation Metrics**:
  - Recognition Accuracy (Acc.)
  - Fréchet Inception Distance (FID)
  - Variation of motion across all words (Div.)
  - Per-word motion variation (Mul.)

- **Qualitative Results**



- **Future Work**:
  - Larger Available Dataset
  - Detailed Facial Expression
  - Open-domain Generation