SignAvatar: Sign Language 3D Motion Reconstruction and Generation

Lu Dong, Lipisha Chaudhary, Fei Xu, Xiao Wang, Mason Lary, Ifeoma Nwogu

University at Buffalo
The State University of New York

Project Page

Motivation

- **Background:** Achieving expressive 3D motion reconstruction and automatic generation for isolated sign words involves three main challenges:
  - Lack of 3D sign-word data.
  - Complex nuances of signing motions.
  - Cross-modal understanding of sign language semantics.
- **Solution:** Introduce a framework, SignAvatar, that is capable of both word-level sign language reconstruction and generation.
  - Utilizes a transformer-based Conditional Variational Autoencoder (CVAE) architecture, leveraging Contrastive Language-Image Pre-training (CLIP) for conditioning.
  - Incorporates a curriculum learning strategy to enhance robustness.
  - Contributes the ASL3DWord dataset and evaluates the proposed framework through extensive experiments.

Semantic: drink

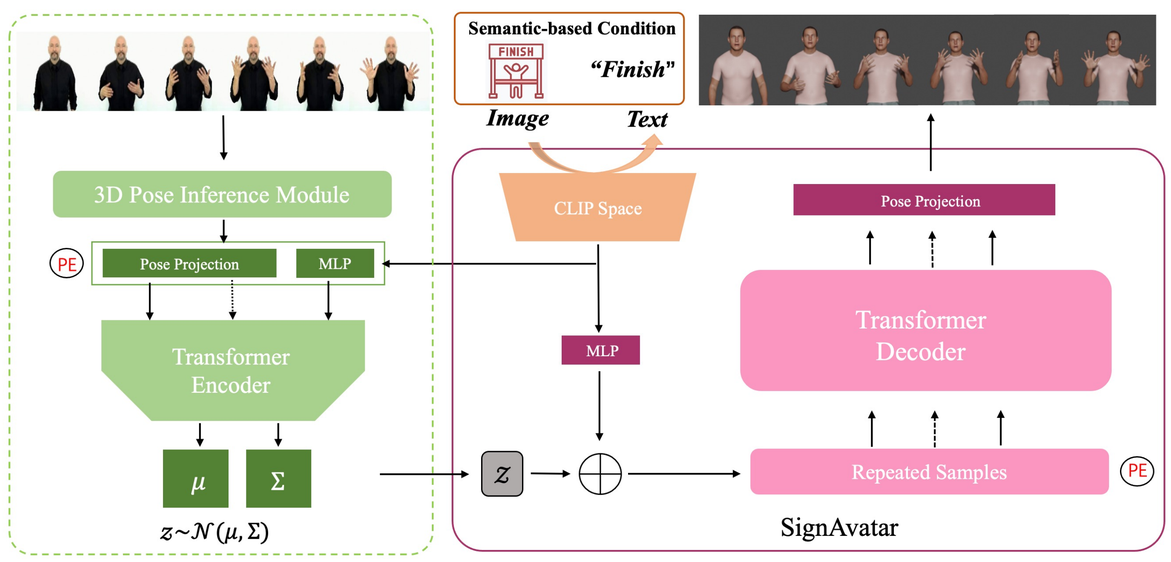
SignAvatar Overview

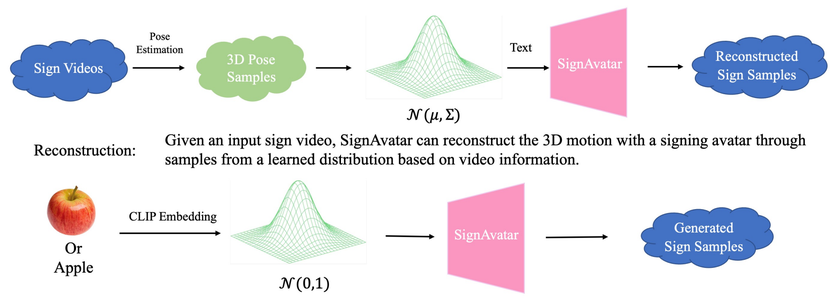
- **Problem Formulation:** Given the input video frames, we reconstruct 3D motion, M_n= [p_1, ...p_T], Meanwhile, given a label Y or an image reflecting that label, we generate the corresponding motion M_n.
- **Conditional VAE with CLIP latent space**
  - ***Encoder:*** The CVAE encoder takes the pose sequence and word-level text projections as input, using a transformer architecture to calculate Gaussian distribution parameters, μ and Σ.
  - ***Decoder:*** Given a latent vector z, with a a conditional bias that integrate categorical information, the decoder generates pose sequence in the SMPL-X parameter format.
  - ***Learning Objectives***

$$\mathcal{L}_{\text{rec}} = \frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_t\|_2^2 \qquad \mathcal{L}_{\text{CVAE}} = \mathcal{L}_{\text{rec}} + \omega_{\text{KL}}\mathcal{L}_{\text{KL}}$$

- **Curriculum Learning Strategy:** This work employs a curriculum learning strategy by progressively increasing the mask ratio during training through $g(ep)$.

$$g(ep) = min\left\{0.1 * \left\lfloor \frac{ep}{500} \right\rfloor, 0.6\right\}, ep \in [0, 5000)$$





Experiments

- **Dataset:** To quantitatively evaluate the performance of SignAvatar, we constructed the ASL3DWord dataset from the WLASL video dataset. However, the distribution of videos for each word is unbalanced and contains a considerable number of noisy samples.
  - Quality Control
  - Pose Feature Extraction
- **Evaluation Metrics:**
  - Recognition Accuracy (Acc.)
  - Fréchet Inception Distance (FID)
  - Diversity (Div.)

$$Diversity = \frac{1}{S_d}\sum_{i=1}^{D_d}\left\|m_i - m_i^{'}\right\|_2$$

  - Multimoldality (Multi.)

$$Multimodality = \frac{1}{C \times S_l}\sum_{c=1}^{C}\sum_{i=1}^{S_l}\left\|m_{c,i} - m_{c,i}^{'}\right\|_2$$

- **Quantitative Results**

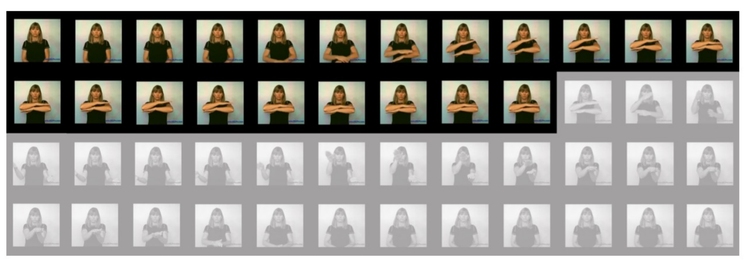
TABLE I: Quantitative results comparison on Raw Poses (*Raw*), Reconstructed Poses (*Rec*), and Generated Poses (*Gen*). → indicates results are better if they are closer to the extracted Raw pose.

ASL3DWord Subset	Acc. ↑	FID ↓	Div.→	Multi.→	ASL3DWord	Acc.↑	FID↓	Div.→	Multi.→
Raw_{train}	1.0	0	30.001	9.921	Raw_{train}	1.0	0	34.565	13.256
Raw_{test}	0.897	0	26.252	11.180	Raw_{test}	0.818	0	30.599	12.289
w/o Curriculum Learning					**w/o Curriculum Learning**				
Rec_{train}	1.0	4.566	28.981	10.160	Rec_{train}	1.0	3.395	33.566	13.803
Rec_{test}	0.962	32.583	29.495	9.095	Rec_{test}	0.906	29.184	31.356	10.249
Gen_{train}	0.884	75.243	24.566	8.250	Gen_{train}	0.515	126.830	25.732	16.500
Gen_{test}	0.890	65.285	24.187	6.600	Gen_{test}	0.5111	100.147	25.393	12.289
w/ Curriculum Learning					**w/ Curriculum Learning**				
Rec_{train}	0.997	7.195	29.340	9.993	Rec_{train}	0.999	6.112	33.583	13.347
Rec_{test}	0.976	32.973	29.165	7.362	Rec_{test}	0.952	40.637	32.561	8.486
Gen_{train}	0.946	44.469	27.115	7.160	Gen_{train}	0.729	85.025	27.973	14.700
Gen_{test}	0.941	46.435	27.097	5.916	Gen_{test}	0.733	71.809	27.811	10.483

TABLE II: Ablation Study for Quality Control

Data	w/o Quality Control		w/ Quality Control	
ASL3DWord Subset	**Acc. ↑**	**FID ↓**	**Acc. ↑**	**FID ↓**
Raw_{train}	1.0	0	1.0	0
Raw_{test}	0.790	0	0.897	0
Rec_{train}	0.927	27.319	0.997	7.195
Rec_{test}	0.851	51.846	0.976	32.973
Gen_{train}	0.856	70.250	0.946	44.469
Gen_{test}	0.860	75.515	0.941	46.435

- **Quality Control**



- **Qualitative Results**